Methuen's Monographs on Biological Subjects

General Editors: G. R. de Beer and Michael Abercrombie

THE MEASUREMENT OF LINKAGE IN HEREDITY

THE MEASUREMENT OF LINKAGE IN HEREDITY

by

K. MATHER, D.Sc., F.R.S.

PROFESSOR OF GENETICS
IN THE
UNIVERSITY OF BIRMINGHAM

AGRICULTURAL COLLEGE AND RESEARCH INSTITUTE
LIBRARY
13 MAY 1952
COIMBATORE

LONDON: METHUEN & CO. LTD.
NEW YORK: JOHN WILEY & SONS, INC.

First published . . April 28th, 1938
Second edition, revised and enlarged 1951

COLLEGE AND RESEARCH INSTITUTE
1? MAY 1952
51063

CATALOGUE NO. 4104/U (METHUEN)

PRINTED IN GREAT BRITAIN

PREFACE

GREGOR MENDEL did not use probability methods in the analysis of his classical experiments, but after the rediscovery of his work, at the beginning of this century, the application of such a technique was soon recognized as a necessary part of genetical analysis. The problems at first were chiefly involved with the testing of the significance of departures from expected ratios, but, with the discovery of linkage, questions of estimation became equally important. The absence of a suitable statistical technique is one of the prime causes of the failure to understand linkage and linkage groups during the years before the Drosophila technique of backcrossing was extensively practised, and the true relations of coupling and repulsion realized. Since that time statistical methods have become more and more favoured by genetical workers and are to-day more than ever necessary for genetical analysis.

Now statistics, like genetics, is a growing science and the crude and often arbitrary methods of yesterday have been superseded and rendered redundant by the development of more exact and more adaptable techniques. This is largely a result of the work of R. A. Fisher and his associates. In some branches of biology, as for example agronomical experimentation, such re-

fined statistical methods are now in common use for several reasons, not the least important of which is the existence of articles and textbooks giving details of their application. In genetics these methods have been less fully described and applied. It is hoped that the present work will serve to bring to the notice of geneticists the desirability of employing such methods and provide the necessary instruction for their use.

It is not claimed that this book is complete. In fact some statistical devices such as the analysis of variance and the use of regressions have been entirely omitted as they are, at present, of secondary importance to the geneticist. On the other hand, the uses of χ^2 and of maximum likelihood have been dealt with in considerable detail as they are of wide application in this branch of science. The methods are described in general and some specific applications are discussed in detail. No attempt has been made to prove all the general formulae used, though some have been considered in detail. Such a course is beyond the scope of this work and would, in any case, merely result in confusion for a non-mathematical reader. If desired, proofs may be found in the original literature cited. But numerical examples have been used wherever possible in the hope that the reader, by working through them, will familiarize himself with the methods and be able to apply them to other problems. The list of general formulae in the last chapter should make reference to any method, and its use for other analyses, a simple matter. It cannot be overemphasized that in order to

make full use of the book the actual examples used should be worked through in detail, as it is only by this means that they will be fully understood and appreciated.

For further details of the genetical and cytological principles, which have been assumed here in order to permit fuller development of the statistics, the reader is referred to *Mendelism and Evolution*, by E. B. Ford and to *The Chromosomes*, by M. J. D. White, both of which are in this series of Biological Monographs.

I am indebted to Professor R. A. Fisher, Head of the Galton Laboratory, for permitting the reproduction herein of a number of tables mainly from the *Annals of Eugenics*, and for continued encouragement during the writing of this book, and to Messrs. W. J. C. Lawrence and J. C. Cullen, of the John Innes Horticultural Institution, for reading and criticizing the manuscript. I also wish to thank Messrs. Oliver and Boyd for allowing me to reproduce Tables I and II, from R. A. Fisher's *Statistical Methods for Research Workers*.

PREFACE TO THE SECOND EDITION

THE chief alterations made in preparing the second edition are the inclusion of a new section in Chapter VII on the scoring method devised by Professor R. A. Fisher for use in testing and estimating linkage, and a chapter on the estimation of gene frequencies, as is needed in the analysis of the distribution of, for example, blood groups in the various races of man. References to certain other developments of the past

twelve years have also been added, and the occasion has been taken to correct a number of minor, though misleading, mistakes. I am indebted to several correspondents for drawing my attention to these errors.

CONTENTS

CHAP.		PAGE
	PREFACE	v
I	INTRODUCTORY	1
II	TWO CLASS SEGREGATIONS	13
III	THE PLANNING OF EXPERIMENTS (I)	26
IV	THE DETECTION OF LINKAGE	32
V	THE ESTIMATION OF LINKAGE	44
VI	INFORMATION AND THE PLANNING OF EXPERIMENTS (II)	56
VII	COMBINED ESTIMATION AND TESTING HETEROGENEITY	69
VIII	DISTURBED SEGREGATIONS	91
IX	HUMAN GENETICS (I)	104
X	HUMAN GENETICS (II)	114
XI	THE ESTIMATION OF GENE FREQUENCIES	125
XII	SYMBOLS AND FORMULAE	137
	REFERENCES	145
	INDEX	147

CHAPTER I

INTRODUCTORY

1. GENETICAL

THE science of Genetics is based on the experiments of Gregor Mendel. His hypothesis of particulate inheritance is the foundation of modern genetical theory and his experimental technique of observing the occurrence and extent of segregations for single factor differences is the basis of modern genetical method.

The field of research covered by Genetics to-day is very large and very varied. The studies of chromosome behaviour, heterozygosity of wild populations and the physiology of gene action, to name but three examples, are concerned with largely different problems and involve largely different techniques. All are, however, alike in involving the study of gene differences. Without the observation and analysis of segregations none of these fields would be open, although other and non-genetical techniques, cytological, anatomical or physiological, could perhaps be used. The genetical method of investigation breaks down unless suitable genes, each comprising two allelomorphs, can be found, and their segregations observed.

Each line of work has commenced from single factor segregations and has, in developing, uncovered more complex interactions and dependencies of the single genes in inheritance. These more complex mechanisms, once understood, have provided tools

for the analysis of still further and still more recondite situations. For example, the detection of dependent segregation of two or more genes has led to the analysis of the organization of linkage groups. The use of linkage as a research tool has in its turn permitted the development of genetical studies on crossing-over and now shows promise of being a powerful agent in the analysis of the complexities of heritable quantitative differences.

Thus any piece of genetical research, in being based on single factor segregations, requires initially a consideration and analysis of the single genes concerned. It is necessary to isolate and identify the genes which are segregating in the material. In some cases, e.g. *Drosophila melanogaster*, *Primula sinensis*, the genes may be well known from past work. This is, however, not always the case and often the primary analysis essential to the future of the research is that of the genes involved.

There are, at least superficially, two different methods of testing the hypothesis that a given distinction between two types is controlled by a single factor. The first is to show that, in a diploid, only three genotypes exist for this factor, that two of them are pure breeding and that only one shows segregation at gametogenesis for the difference between the two postulated allelomorphs. Segregation is the occurrence of two kinds of gametes distinguishable by their capacity for producing distinct genetical types (e.g. the production of **AA** as opposed to **Aa** or of **Aa** as opposed to **aa**) when mated with any given single type of gamete. The differences observed in the progeny of such an individual whose gametes show segregation are also referred to as segregations.

Tests of this first type are perhaps the most convincing, but are sometimes not immediately available, and, in any case, are often preceded as evidence by the second type of data, viz. that of the numerical

relations of the classes in a segregating progeny. In the case of a single factor there exists but one type, the *heterozygous* or **Aa** type, capable of producing two different (**A** and **a**) gametes. Furthermore, the hypothesis states that it will produce such gametes in equal numbers. Then on *backcrossing* such a heterozygote to a pure recessive, i.e. **Aa** × **aa**, a 1 : 1 segregation for types **Aa** : **aa** should be obtained. On intercrossing two heterozygotes an F_2 ratio of 3 : 1 for **AA** + **Aa** : **aa**, or 1 : 2 : 1 for **AA** : **Aa** : **aa** in the absence of dominance, should be found. These are the only segregations possible for an uncomplicated single factor difference. We can now ask the question, 'Are the observed ratios in agreement with these expectations?'

If the segregations are tested by tetrad analysis, i.e. the testing of all four gametes which are the products of any meiotic division, such as is possible in some lower plants, exact 1 : 1 gametic segregation should be observed. This is, however, not always the case in the progeny of heterozygotes, as normally obtained. Chance deviations from expectation can occur. Then the use of this method of testing the single factor hypothesis automatically involves consideration of chance deviations in the segregations.

Other hypotheses, e.g. of two complementary factors or of a single gene with one *homozygous* (**AA** or **aa**) form lethal, can be tested in both of these ways. It can be shown that more than two pure breeding types exist in the one case and that but one such type exists in the other. It can also be shown that an F_2 ratio of 9 : 7 and a backcross ratio of 1 : 3 can be obtained for two complementary factors and that the lethal single gene gives 2 : 1 and 1 : 1 segregations in F_2 and backcross respectively. Precisely the same principles are involved as in the testing of the hypothesis of a simple single gene.

The other situation often met with and also often important in genetical analysis is that of linkage. In some families it may be necessary to distinguish linkage and factor interaction, or it may be necessary to have exact knowledge of the linkage relations of two genes prior to their use in the analysis of quantitative characters, or the information may be necessary for a number of other reasons.

Now, linkage can only be observed in families segregating for each of the two characters observed, and, furthermore, only if at least one parent is doubly heterozygous. For example, a cross of the type **Aabb** × **aaBb** gives no information about linkage. If the two genes are not linked the gametic segregation of the double heterozygote will be $\frac{1}{4}$ **AB**, $\frac{1}{4}$ **Ab**, $\frac{1}{4}$ **aB**, $\frac{1}{4}$ **ab**, and this may be realized in tetrad analysis. Among the progeny of a cross it may not be found exactly, because chance deviations will again occur.

If linkage exists between the two genes, the gametic output of the double heterozygote is $\frac{1}{2}(1-p)$ **AB**, $\frac{1}{2}p$ **Ab**, $\frac{1}{2}p$ **aB**, $\frac{1}{2}(1-p)$ **ab**, or $\frac{1}{2}p$ **AB**, $\frac{1}{2}(1-p)$ **Ab**, $\frac{1}{2}(1-p)$ **aB**, $\frac{1}{2}p$ **ab**, where p is the *recombination fraction* and has the value 0·5 when there is no linkage. If after considering chance deviations and other complications p is demonstrably different from 0·5 the evidence for linkage is clear.

There is another test for the presence of linkage. In a double heterozygote two relations in arrangement may exist between the genes. Either **A** — **B** and **a** — **b** may be on the same chromosomes or **A** — **b** and **a** — **B** may be the arrangement found. In the former case **A** — **b** and **a** — **B** are the recombination types. In the latter they are **A** — **B** and **a** — **b**. If two different arrangements, bearing these relations, can be shown to exist, linkage is demonstrated. These arrangements are differentiated under the names of *coupling* and *repulsion*, the arrangement to which a particular name is allocated

being dependent on conventions varying with different organisms and different schools of research. Usually, however, coupling is the $\frac{\mathbf{AB}}{\mathbf{ab}}$ type and repulsion the $\frac{\mathbf{Ab}}{\mathbf{aB}}$ type.

When linkage is once detected, by its causing characteristic deviations from the 1 : 1 : 1 : 1 expected in the double backcross (**AaBb** × **aabb**) or the 3 : 1 : 3 : 1 expected from the single backcross (**AaBb** × **aaBb** or **AaBb** × **Aabb**) or the 9 : 3 : 3 : 1 of the F_2 (**AaBb** × **AaBb**), it is usually necessary to measure it. The common measure, developed from the chromosome theory of heredity, is by the calculation of the recombination fraction or value, denoted above by p. The value is a measure, though not always a simple one, of the frequency of crossing-over between the two chromosomes in the region delimited by the two genes under consideration.

It is in the consideration of the chance variations from expectation, such as has been shown to occur at almost every stage of the genetical mechanism, that statistical methods are necessary.

2. STATISTICAL

It is clear that the 3 : 1 and 1 : 1 segregations of single factor differences will seldom be realized exactly, because the individuals of the progeny represent samples from a large population of gametes, half of which carry one allelomorph and half the other. Similarly, recombination values of 50 per cent will not be exactly realized even though the genes are carried by different chromosomes, because, again, the progeny represents a series of samples from a population of gametes of which half are recombinations. The differentiation of such chance fluctuations from real deviations requires a *test of significance* of the observed departure from the expected ratio.

The principle underlying such a test of significance is simple but must be grasped clearly. The results observed are compared with those expected on the basis of the hypothesis under consideration. The probability of obtaining by chance a departure from expectation at least as large as that found is calculated, and if this probability is sufficiently small it is concluded that the departure is significant. What constitutes 'sufficiently small' is dependent on circumstances. If a single family segregates in such a way that its departure from expectation would be equalled or exceeded by chance in but one trial out of twenty, it is usually considered to be showing a significant deviation from the hypothesis. But if one family out of twenty was showing such a deviation it could not be considered as indicating significant deviation, because one family out of twenty is expected to do so by chance. The second case differs from the first in that we have had twenty trials before finding a deviation of this magnitude. In such a case it is expected. In the first case of only one family it would be a relatively remote contingency. The test of significance must be capable of dealing with such contingencies as these.

An hypothesis can never be proved or disproved by a test of significance. If the data do not show a significant deviation from expectation they agree with the hypothesis, but they may also agree with several other hypotheses giving closely similar expectations. The simplest or most relevant hypothesis is considered and is not discarded if the data agree with it, irrespective of how many more complicated hypotheses are also in agreement with observation.

If the data show a high deviation from the expected segregation they do not generally disprove the hypothesis; they only make it a more or less unlikely one. In the case considered above, when only one family was grown, a deviation which would be

exceeded or equalled once in twenty trials was found. The hypothesis is then rendered unlikely as it could account for such a family only once out of twenty times. When twenty families are grown it can account for one such family in each trial and is not unlikely. The level of probability chosen as indicating significant departure from hypothesis is simply the level at which the worker is willing to be misled. If, as is usual, the one in twenty level is taken, he will find that his supposedly real departures are actually chance ones, once in twenty cases. If the one in a hundred level is taken he will be wrong, in calling the departure real, less often, but if a hundred such cases are taken he must expect to be wrong once. If this is constantly borne in mind the experimenter will set his levels of significance to suit his circumstances and will not be disconcerted when an apparently promising line of work comes to nothing because it was based on a false conclusion as a result of his test of significance misleading him.

The only exception to this rule, in genetics, is when segregation is observed to occur in what should be, by the hypothesis, a homozygote. Even this cannot be considered as a complete exception, as the 'segregation' could be the result of mutation, or error in the handling of the material.

When the first and simplest hypothesis has been shown to be unlikely, another may be set up and tested. The new hypothesis may involve a *parameter*, a numerical quantity characterizing the population, which must be estimated, e.g. the supposition that two genes are linked demands that an estimate of the recombination value be obtained before the hypothesis of linkage is sufficiently precise to be tested by observation. This involves the use of a *method of estimation.* When the hypothesis has been formulated precisely it may, in its turn, be tested against the observational data by a new test of significance.

Both tests of significance and methods of estimation rest on considerations of *frequency distributions.* A frequency distribution gives the relative frequencies with which certain events will occur, or individuals be found to fall into certain classes. As an example let us consider the segregation in a single factor backcross.

Any individual in a family showing a backcross segregation is equally likely to have arisen from the union of an **A** gamete or of an **a** gamete from the heterozygous parent, with one of the gametes (all **a**) from the recessive parent. We may then represent the frequencies of one individual falling into the classes **Aa** and **aa** as $\frac{1}{2} : \frac{1}{2}$. A second individual is also equally likely to be **Aa** or **aa** and its character will be independent of that of the first one. Then both will be **Aa** in $\frac{1}{2} \times \frac{1}{2}$ of cases and both will be **aa** in $\frac{1}{2} \times \frac{1}{2}$ of cases. One will be **Aa** and the other **aa** in $2 \times \frac{1}{2} \times \frac{1}{2}$ of cases as this type of family may occur in either of two ways, viz. the first individual may be **Aa** and the second **aa** or the second **Aa** and the first **aa**. The frequencies with which the three types of family will occur are thus :—both **Aa** $\frac{1}{4}$, one **Aa** and one **aa** $\frac{1}{2}$, and both **aa** $\frac{1}{4}$.

A similar argument leads to the conclusion that families of three individuals will show 3, 2, 1, and 0 **Aa** individuals in 1/8, 3/8, 3/8, and 1/8 of cases respectively.

It will be observed that these frequencies for families of one, two and three individuals are given by the expansions of the binomial expressions $(\frac{1}{2} + \frac{1}{2})^1$, $(\frac{1}{2} + \frac{1}{2})^2$ and $(\frac{1}{2} + \frac{1}{2})^3$ respectively. The general form for a family of n individuals is $(\frac{1}{2} + \frac{1}{2})^n$. We can calculate the expected frequencies of the various types of family of size n by expanding this formula.

Suppose that we have a family of eight individuals expected to be segregating in the ratio 1 : 1. The frequencies of families with 8, 7, 6, 5, 4, 3, 2, 1, and

0 **Aa** individuals will be, from the expansion of $(\frac{1}{2}+\frac{1}{2})^8$:

No. of Aa individuals	8	7	6	5	4	3	2	1	0
Frequency . .	$\frac{1}{256}$	$\frac{8}{256}$	$\frac{28}{256}$	$\frac{56}{256}$	$\frac{70}{256}$	$\frac{56}{256}$	$\frac{28}{256}$	$\frac{8}{256}$	$\frac{1}{256}$

If we observe that a family actually obtained shows 7 **Aa** and 1 **aa** members, we can apply a test of significance of the deviation from the expected 4 : 4 by the use of the above frequency distribution. The deviation is 3 from expectation in each class. Now deviations of 3 or more will occur when families of 8 : 0, 7 : 1, 1 : 7, or 0 : 8 are found. These families are expected with the frequencies $\frac{1}{256}$, $\frac{8}{256}$, $\frac{8}{256}$, and $\frac{1}{256}$, respectively. Then the total expectation of obtaining as large or larger deviation than the one observed is $\frac{1+8+8+1}{256}$ i.e. $\frac{18}{256}$ or 0·070. This is slightly greater than 7 per cent. It is generally considered that a deviation should not be taken as significant until the probability of obtaining it by chance is less than 5 per cent, and so the hypothesis can, in this case, be accepted as agreeing sufficiently well with the data.

If we had been expecting a segregation of 3 : 1, or $\frac{3}{4}:\frac{1}{4}$, the correct binomial for a family of n would have been $(\frac{3}{4}+\frac{1}{4})^n$. The general form for a ratio of $x : y$, where $y = 1 - x$, is $(x + y)^n$ and the general term of the expansion giving the frequency of families with r individuals of one class and $n - r$ of the other is

$$\frac{n!}{r!(n-r)!}(x)^r(y)^{n-r}$$

While it is always possible to do a test of significance in this way, it is not always convenient to calculate the binomial expansion, particularly when n is large. A quicker technique is needed. Now the probability of a deviation being equalled or exceeded by chance is expressible as a function of the deviation divided

by a quantity called the *standard error.* This quantity, denoted by σ, is really a measure of the spread of the frequency distribution and is calculated from the formula $\sigma = \sqrt{xyn}$.[1] The probability of the deviation being equalled or exceeded is tabulated against the deviation : standard-error ratio for ease of use. It can be calculated for any particular example, but is more readily available in tabular form. The important point about this method is that for large samples the probability corresponding to any given ratio of deviation and standard error is constant no matter what the mean and standard deviation of the distribution may be. The limitation of the method is that it assumes a continuous distribution, whereas the binomial is really discontinuous. On the other hand, as n increases the discontinuity becomes less and less important, and may be neglected for quite low values of n. Even where discontinuity does seriously affect the result it may be corrected quite easily, as will be seen later. The standard error technique is based on the use of the *normal distribution* which is the limit reached by the binomial distribution when n is infinite. This technique has been very popular with geneticists in the past ; the ratio of deviation to standard error being used under the symbol of d/m.

Other quantities calculated from the deviation and the expectation can be used in the same way, when their relations to the probability have been determined and tabulated. One in particular, χ^2, is of great value as it is additive, i.e. the sum of two

[1] It is customary to denote a parameter by a Greek letter and the corresponding statistic, or estimate of the parameter, by a corresponding Latin letter. Thus the standard error of the binomial expansion $(x + y)^n$, where x and y are fixed by hypothesis is not an estimate and is denoted by σ. The standard error of this expansion if x were estimated would be itself an estimate of the true standard error σ and should be denoted by s. This convention will be followed with all the symbols used.

independent χ^2 quantities is itself a χ^2 and may be used as a joint test of significance.

Methods of estimation are also related to the binomial expansion. This expansion is itself a special case of the multinomial $(m_1 + m_2 + m_3 \ldots)^n$ whose general term is

$$\frac{n!}{a_1!a_2!a_3! \ldots}(m_1)^{a_1}(m_2)^{a_2}(m_3)^{a_3} \ldots$$

where $m_1 + m_2 + m_3$, &c., is 1 and $a_1 + a_2 + a_3$, &c., is n.

The *method of maximum likelihood* which has the property, unique among methods of estimation, of always extracting the most precise estimate which the data can yield, is based on this multinomial expansion. Just as we could express the chance of finding, in a family of n, r of one type and $n - r$ of the other, the expectation of any individual falling in the first class being x, as $\frac{n!}{r!(n-r)!}(x)^r(y)^{n-r}$ so we can express the chance or likelihood of finding a family of n individuals with a_1 of the first kind, a_2 of the second, and so on, by the general term of the multinomial above. Then when m_1, m_2, &c., are known in terms of the parameter we wish to estimate, e.g. when they are the expectations of the four classes of a backcross for two linked factors expressed as a function of p the recombination value, this term of the multinomial is the chance or likelihood of finding such a family, expressed in terms of what we want to measure. That value of the variable which makes this likelihood a maximum, is then found and is taken as the best estimate of the parameter.

Now the estimate of a parameter, or *statistic* as it is termed, derived by some such process, will deviate from the true value of the parameter as a result of sampling variation, just as families expected to give a 1 : 1 ratio deviate from this ratio by sampling error. For example, if in a backcross we found 15

of one class and 20 of the other, an estimate of the frequency of gametes which give rise to the first kind of individual would be $\frac{3}{7}$. This is not $\frac{1}{2}$, but it must be considered as an estimate of $\frac{1}{2}$, because the family is in keeping with the backcross expectation, the deviation being ascribable to pure sampling error. Thus we require some measure of the confidence which can be reposed in an estimate of some parameter. This is given by the standard error of the estimate, or very often by the *variance*, which is the square of the standard error. The variance and standard error are measures of the spread of the distribution of the estimate round its true value, the parameter, and so are measures of the precision with which the estimate is made. The method of maximum likelihood always has the maximum precision possible as measured by this means.

These principles, outlined above for simple cases, are the bases of all methods of analysing genetical data. The frequency distribution, or such statistics as specify it, of some quantity or quantities are calculated and are used in the test of the hypothesis. Complications may be introduced by complexities or shortcomings of the data, and the analysis may need to be complicated to accommodate such data, but the essentials of the methods remain the same.

CHAPTER II

TWO CLASS SEGREGATIONS

3. DEVIATION AND HETEROGENEITY

ANALYSIS of the single factor segregations is the first consideration not only because of the interest which may attach to them themselves, but also because the subsequent treatment of the data will depend to some extent on the nature of these single factor ratios.

Two questions may be asked about the segregation of a single factor: (*a*) Is it in keeping with some expected ratio, e.g. 3 : 1 or 1 : 1 ? (*b*) Are all the families in agreement in showing the same result, i.e. are the data homogeneous ? The answers to both questions are provided by suitable tests of significance.

4. THE USE OF THE STANDARD ERROR

One of the most popular tests of significance used in detecting deviations from expected ratios is that based on the standard error, discussed in the previous chapter. The standard error of a binomial distribution, $(x + y)^n$ is given by the formula

$$\sigma_x = \sqrt{\frac{xy}{n}} \text{ where } x + y = 1$$

This is the standard error appropriate to testing the agreement of that observed proportion of the family which falls into one class with its expected value

x or y. If we want to test the agreement of the actual number of individuals in one class with the number expected we use the standard error

$$\sigma_{xn} = n\sqrt{\frac{xy}{n}} = \sqrt{xyn}$$

The procedure, as noted before, is simple. The standard error is found, whether for proportion or number, and the corresponding deviation from expectation is found for either the proportion or number observed to be in one class. The deviation is divided by the standard error and then by the use of a ‘Table of Normal Deviates’, such as is provided at the end of this book (Table I), the corresponding probability of obtaining as large or larger deviation by chance, is obtained. The procedure may be illustrated by an example.

Ex. 1. In a family of *Antirrhinum majus*, obtained by selfing a yellow-flowered plant known to be heterozygous, the following segregation for flower colour was observed.

Yellow-flowered plants	208
Ivory	81
Total in family	289

Is this in keeping with the 3 : 1 ratio expected from selfing a plant heterozygous for a single flower colour gene ?

When a 3 : 1 ratio is expected the frequencies of families of 289 individuals, with 289, 288, &c. yellow plants will be given by the expansion of $(\frac{3}{4} + \frac{1}{4})^{289}$.

The number expected in each of the two classes will be $\frac{3}{4} \times 289$ and $\frac{1}{4} \times 289$ respectively, i.e. 216·75 yellows and 72·25 ivories. Then the deviation of the number observed in each class from expectation is 216·75 – 208, i.e. 8·75. The standard error of the number in each class is $\sqrt{\frac{3}{4} \times \frac{1}{4} \times 289}$, i.e. 7·36.

Hence the ratio of the deviation to the standard error is

$$\frac{d}{\sigma} = \frac{8{\cdot}75}{7{\cdot}36} = 1{\cdot}19$$

Reference to Table I shows that such a deviation, 1·19 times the standard error in size, is expected by chance about once in four trials, or, in other words, has a probability of about 0·23. If this difference were to be considered as indicating real deviation from the hypothesis, then a false conclusion would have been reached once in every four cases. This is too great a proportion of errors, and so the data must be considered as agreeing sufficiently well with the hypothesis. In general, no probability greater than 0·05, i.e. one in twenty, should be considered as indicating significant deviation.

The standard error is of use and is easy to apply to such cases as the above where only one family is concerned, and this family segregates into but two classes. The standard error is, however, not easy to use for testing for deviations from hypotheses when more classes than two are observed, nor is it easily adaptable to the testing of agreement among several families. For these purposes Pearson's χ^2 is much to be preferred.

5. TESTING DEVIATIONS BY χ^2

χ^2 is calculated from the general formula

$$\chi^2 = S\left[\frac{(a - mn)^2}{mn}\right]$$

where a is the number observed and m the proportion expected in a class, n is the total and S stands for summation over all classes. This quantity is just as simple to use for testing deviations in a single family segregating into but two classes, as is the standard error.

Ex. 2. We may illustrate its simple use in this way, by considering the Antirrhinum data quoted in the last example. The setting out of the data and

the calculation of χ^2 testing the deviation from the expected 3 : 1 is shown in Table 1.

TABLE 1

Class	Number Observed (a)	Number Expected (mn)	Deviation ($a - mn$)	$\chi^2\left(\frac{(a-mn)^2}{mn}\right)$
Yellow . .	208	216·75	− 8·75	0·353
Ivory . .	81	72·25	+ 8·75	1·056
Total . .	289	289·00	0·00	1·409

There is one further determination to make before χ^2 can be entered in the corresponding table of probabilities, viz. the number of '*degrees of freedom*' to which χ^2 corresponds. The rule for determining the number of degrees of freedom is simply stated as 'the number of degrees of freedom is the number of classes which can be filled arbitrarily'. This and subsequent examples will amply illustrate the use of this rule. In the present case only one class could be filled arbitrarily, for once a number were assigned to the yellow class the number in the ivory class would follow, because it is the total minus the number of yellows. We have, then, one degree of freedom.

The table of probabilities (Table II) given at the end of this chapter is taken from Fisher (1949*a*). The probability of obtaining as large or larger deviation is given at the head of the table, and each row of the table corresponds to a number of degrees of freedom as shown in the leftmost column. The body of the table contains the χ^2 values. To use the table we note that we have one degree of freedom, and so must use the first row. Then our value of χ^2, 1·409, lies between those values whose probabilities are 0·3 and 0·2. This is in agreement with the standard error test, as indeed it must be if the tests are both suitable and both calculated correctly.[1] It is unnecessary to know the probability with any further

[1] For one degree of freedom $\chi^2 = \left(\frac{d}{\sigma}\right)^2$.

accuracy as it cannot be considered as indicating a significant deviation unless as low as 0·05.

6. χ^2 AS A TEST OF HETEROGENEITY

The above simple example fails to bring out the really valuable property of χ^2, which is its additive character. The sum of two χ^2s is itself a χ^2 for a number of degrees of freedom obtained by adding the two numbers of degrees of freedom corresponding to the initial χ^2s. It is this additive property which allows of the easy testing of homogeneity, as illustrated by the next example.

Ex. 3. Fisher and Mather (1936) give the results of a backcross for several factors in mice. Table 2 shows the segregations observed for the genes **D,d** (intense-dilute coat colour) and **Wv,wv** (straight-wavy hair) in the five groups into which this backcross is divided. The totals of the intense and dilute, and straight and wavy animals for the whole backcross are shown at the bottom. In each case there is a shortage of recessives from the expected half. Are these shortages of recessives significant and are the families homogeneous ?

TABLE 2

SEGREGATIONS FOR THE FACTORS **D,d** AND **Wv,wv** IN A MOUSE BACKCROSS

Group	D	d	χ^2	Wv	wv	χ^2
1	219	211	0·1488	209	221	0·3349
2	174	137	4·4019	169	152	0·9003
3	96	72	3·4286	91	82	0·4682
4	31	28	0·1525	36	23	2·8644
5	128	123	0·0996	134	117	1·1514
			8·2314			5·7192
	648	571	4·8638	639	595	1·5689

	χ^2	D.f.	P	χ^2	D.f.	P
Deviation .	4·864	1	0·05 – 0·02	1·569	1	0·3 – 0·2
Heterogeneity	3·367	4	0·5	4·150	4	0·5 – 0·3
Total . . .	8·231	5		5·719	5	

Let us first consider the **Wv,wv** segregation. χ^2 for the deviation from the expected 1 : 1 segregation is calculated by the method of the previous example, for each group separately. For example, the first group of 430 mice is expected to show 215 **Wv** and 215 **wv** mice. It actually has 209 **Wv** and 221 **wv**. Then

$$\chi^2 = \frac{(215 - 209)^2}{215} + \frac{(215 - 221)^2}{215} = 0{\cdot}3349$$

Each group yields a χ^2 for one degree of freedom. Hence the sum of these five χ^2s, 5·7192, is itself a χ^2 for five degrees of freedom. This total χ^2 may be considered as comprising two parts, (*a*) a portion concerned with the grand deviation of all the groups taken together from the expected 1 : 1, and (*b*) a portion concerned with the disagreement among the groups when allowance has been made for the grand deviation. Now the former portion may be calculated from the totals of straight and wavy mice in all the groups taken together. It is found to be 1·5689 by precisely the same type of calculation as for the single groups, and will have one degree of freedom because it is concerned with a distinction into two classes. The difference of the total χ^2 for five degrees of freedom and this χ^2 for one degree of freedom, calculated from the combined segregation, will be the second or heterogeneity χ^2 testing the agreement between the five groups. This heterogeneity χ^2 must have $5 - 1$, i.e. four degrees of freedom. Thus we get the analysis of χ^2 into its two parts, testing deviation from 1 : 1 and heterogeneity among groups respectively, as shown below Table 2. Neither χ^2 when referred to Table II has a significantly low probability, and so we may say that the groups agree (*a*) with the expected 1 : 1 segregation and (*b*) with one another. The latter agreement considerably increases confidence in the value of the former agreement.

The data for **D,d** are analysed in a precisely similar manner. In this case, however, the 'deviation' χ^2, calculated from the total segregation, is 4·864 for one degree of freedom. This χ^2 is found from Table II to have a probability of between 0·05 and 0·02. Such a large deviation could be obtained by chance less than once in twenty trials, and so should be considered as indicating a real shortage of **dd** mice. The 'heterogeneity' χ^2, obtained as before, by subtraction is 8·231 − 4·864, i.e. 3·367 for four degrees of freedom and has a probability of 0·5. The groups thus agree with one another in showing a shortage of **dd** individuals.

The agreement among the groups settles any doubts as to the reality of the **dd** shortage. It removes the suspicion that the shortage is due to faulty experimental technique.

It may be noted here that, in general, heterogeneity, if established, is often a direct result of faulty experimentation. It may flow from poor classification or partial selection of stronger types by overcrowding or insufficient feeding, and from other *similar causes. With inexplicably heterogeneous* data the whole experiment and its technique is suspect. With absence of significant heterogeneity, as in the above example, the validity of the deviations is not called into question.

One further point must be made about the last example. The total segregation of **Dd** : **dd** mice did not agree with the expected 1 : 1, and yet the heterogeneity χ^2 was calculated on the assumption that observation and expectation did not disagree. Hence the heterogeneity χ^2 obtained by subtraction is not absolutely accurate. In the actual practice, however, it needs a considerably greater deviation of total segregation from expectation seriously to invalidate a heterogeneity χ^2 calculated in this way. In this example the true value, calculated from the observed segregation of 648 **Dd** : 571 **dd** is 3·381

whereas the value obtained by assuming a 1 : 1 segregation was 3·367. This difference would cause no serious misjudgement. Where, however, the deviation of the total segregation is more serious, a more trustworthy method of calculating the χ^2 for heterogeneity must be used, in order to avoid the liability of serious misjudgement. Such a case is also provided by Fisher and Mather's mouse backcross.

7. HETEROGENEITY WHEN SIGNIFICANT TOTAL DEVIATIONS ARE PRESENT

Ex. 4. In Table 3 are set out the details of the segregation for **T,t** (dark head—light head) in Fisher and Mather's mouse backcross. It will be seen that the 1,013 mice are divisible into five groups according to the type of the male parent. These types were distinguished, before the commencement of the backcross, by their origin, and the grouping is in no wise dependent on the breeding results. Two of the male types comprise but one individual each, but the other three groups each contain several individuals. We have thus a hierarchical classification, the whole experiment being divisible into five major groups and each major group further subdivisible into smaller subgroups.

There are twenty-one such subgroups and on calculating a χ^2 for the **T,t** segregation in each, and summing the results we should have a total χ^2 for twenty-one degrees of freedom. A χ^2 for five degrees of freedom can also be calculated from the segregation totals of the male type groups. There are five types and each will give a χ^2 for one degree of freedom, hence the total obtained by summing the male type χ^2s will have five degrees of freedom. Finally a χ^2 for one degree of freedom can be calculated from the grand total segregation and will serve to detect deviation from the expected 1 : 1 ratio. Now the heterogeneity between the five male type segregations could be obtained, on the assumption of a 1 : 1

TABLE 3

SEGREGATION FOR THE FACTOR T,t IN A MOUSE BACKCROSS

Individual Males		Classes of Male		Total	
a_1	n	a_2	n	a_{2t}	n_t
60	128				
38	96				
6	21				
4	16	155	359		
27	55				
10	17				
10	26				
37	92				
18	45				
10	33	90	233		
11	21			427	1013
14	42				
49	122	49	122		
27	59	27	59		
37	80				
19	40				
20	49				
3	6	106	240		
3	12				
12	32				
12	21				

segregation, by subtracting from the summed χ^2 of the group segregations, the χ^2 for the total deviation. This operation would be precisely the same as in the last example. The χ^2 for heterogeneity between male types would then have four degrees of freedom. Similarly a χ^2 for heterogeneity between individual males, but corrected for heterogeneity between groups, would be found by summing the twenty-one individual male χ^2s and subtracting from the resulting total the summed χ^2 obtained from the five group segregations. The analysis would thus be :

		D.f.
	Deviation from 1 : 1 . .	1
Heterogeneity	Between male types . .	4
	Between individual males .	16
	Total	21

In Table 3 are given, for each individual male, each male group and the whole experiment, the total mice raised (n) and the number of **t** mice (a_2). There are in all 427 **tt** mice out of a total of 1,013. This is very much less than the expected 506·5, the deviation being significant as measured by χ^2. Thus in computing the heterogeneity χ^2 a 1 : 1 ratio must not be assumed. We must take the observed total segregation of 586 : 427 and calculate the heterogeneity χ^2 on this basis. We shall thus reduce the deviation χ^2 to zero or in other words shall 'lose' one degree of freedom by calculating the best fitting total segregation, i.e. fitting the parameter x in $(x + y)^n$. This principle of losing a degree of freedom on fitting a parameter will be used extensively later.

The calculation of χ^2 could be done as before. The expected numbers of **T** and **t** mice in each family or group would be given by $\frac{586}{1013}n$ **T** and $\frac{427}{1013}n$ **t**, where n is the family or group total, instead of $\frac{1}{2}n$ and $\frac{1}{2}n$ if the 1 : 1 were to be assumed. This is, however, a laborious method and an easier one developed by Brandt and Snedecor (cf. Fisher, 1949*a*) can be used.

For each individual male's family we calculate the quantity $\frac{(a_2)^2}{n}$, where a_2 is the number of **t** mice and n the family total. The same is done for the male type groups and also for the whole experiment. The $\frac{(a_2)^2}{n}$ values are proportional to χ^2 for that family. These $\frac{(a_2)^2}{n}$ values are entered in Table 4 in the same arrangement as the data of Table 3. In Table 4 we have three columns of values (the rightmost having but one entry) and each column is summed. The totals are proportional to the χ^2s corresponding to (left) twenty-one individual males,

TABLE 4

VALUES OF $\frac{(a_2)^2}{n}$ DERIVED FROM TABLE 3

Individual Males	Classes of Male	Total
28·1250		
15·0417		
1·7143		
1·0000	66·9220	
13·2545		
5·8824		
3·8461		
14·8804		
7·2000		
3·0303	34·7639	
5·7619		179·9891
4·6667		
19·6803	19·6803	
12·3559	12·3559	
17·1125		
9·0250		
8·1633		
1·5000	46·8167	
0·7500		
4·5000		
6·8571		
Totals 184·3474	180·5388	179·9891

Differences	3·8086	0·5497
χ^2	15·619	2·254
Degrees of freedom	16 (i.e. 21 − 5)	4 (i.e. 5 − 1)
Probability	0·5 − 0·3	0·7 − 0·5

(middle) the five types of males, and (right) whole experiment, i.e. deviation. Then by taking the difference between the male group total and the whole experiment value (middle and rightmost columns) we obtain a quantity proportional to χ^2 for heterogeneity between the five groups. Similarly the difference between the leftmost and middle totals is proportional to χ^2 for heterogeneity between males of the same group. These differences are converted into χ^2s by multiplying by $\frac{(n_t)^2}{a_{1t}a_{2t}}$ where a_{1t} is the number

of **T** mice, a_{2t} the number of **t** mice and $n_t = a_{1t} + a_{2t}$, in the whole experiment. (This is the adjustment for the bad **T** : **t** segregation. $\frac{(n_t)^2}{a_{1t}a_2}$ would be $\frac{2^2}{1 \times 1}$ if a 1 : 1 ratio were to be assumed.)

This multiplier is $\frac{(1013)^2}{586 \times 427}$ i.e. 4·10103. Then the male type heterogeneity is 0·5497 × 4·10103, i.e. 2·254 for four degrees of freedom and is not significant. The individual male heterogeneity is 3·8086 × 4·10103, i.e. 15·619 for sixteen degrees of freedom and is not significant. Thus there is no heterogeneity and all the families agree in showing a serious shortage of **tt** mice.

If there had been heterogeneity between types of male it would have been necessary to consider the individuals of each group separately from those of other groups. Each group would then have had its own multiplier based on the group values for a_1, a_2, and n. An example illustrating this procedure will be found in Fisher (1949*a*).

8. THE CALCULATION OF χ^2

In the examples worked in this chapter various formulae for calculating χ^2 have been used. There are a number of others, each suited for particular purposes, some of which may be conveniently noted here. Others will be given in later sections. The fundamental formula applicable to all cases is $\chi^2 = S\left[\frac{(a - mn)^2}{mn}\right]$ where a and mn are the observed and expected numbers in any class and S stands for summation over all classes. This formula may be given in a slightly different and more useful form

$$\chi^2 = S\left[\frac{a^2}{mn}\right] - n$$

where a and m are as before and n is the total number

of individuals in the data. This is identical with the previous formula and is also universally applicable.

For a family segregating into two classes, whose expected ratio is $l:1$, and the observed numbers are $a_1 : a_2$

$$\chi^2 = \frac{(a_1 - la_2)^2}{ln}$$

The special cases of this formula for the more usual genetical expectations are :

Ratio	Formula
1 : 1	$\frac{1}{n}(a_1 - a_2)^2$
3 : 1	$\frac{1}{3n}(a_1 - 3a_2)^2$
15 : 1	$\frac{1}{15n}(a_1 - 15a_2)^2$
1 : 3	$\frac{1}{3n}(3a_1 - a_2)^2$
9 : 7	$\frac{7}{9n}\left(a_1 - \frac{9}{7}a_2\right)^2$

The Brandt and Snedecor formula for the calculation of heterogeneity χ^2 from a hierarchical table, like Table 3, is

$$\chi^2 = \frac{(n_t)^2}{a_{1t}a_{2t}}\left[S\left(\frac{a_1{}^2}{n}\right) - \frac{(a_{1t})^2}{n_t}\right]$$

$$= \frac{(n_t)^2}{a_{1t}a_{2t}}\left[S\left(\frac{a_2{}^2}{n}\right) - \frac{(a_{2t})^2}{n_t}\right]$$

CHAPTER III

THE PLANNING OF EXPERIMENTS (I)

9. FAMILY SIZE

IT is usual in genetical work for the scope of the experiments to be limited by such considerations as available space, labour, &c. It is thus necessary to make the best and most profitable use of the numbers of individuals that can be raised. The achievement of this end usually requires considerable care in the planning of the experiments, and statistical methods are often of great value in this connexion.

In many experiments it is desirable to be able to pick out certain genotypes, usually homozygotes, for the purpose of establishing permanent lines or for the detection of some form of factor interaction. This involves making test crosses, the homozygote being distinguished from heterozygous individuals by the failure of segregation in its progeny. These progenies may be of very little value except for this specific purpose and so should be kept as small as is consistent with this end. Now, any progeny failing to segregate may still come from a heterozygous pa[illegible] it as the size of the family increases the chan[illegible] ose which fail to segregate having come fro[illegible] erozygote becomes steadily smaller. The mini[illegible] ze of the progeny designed to test some indi[illegible] then a statistical question involving considera[illegible] he probability that any individual, in a fam[illegible] ed from a heterozygote, will be of the reces[illegible], and also of the permissible maxi-

mum probability of obtaining a misleading result, as decided on by the experimenter.

Let us consider an example. Suppose it is desired to test a series of individuals phenotypically dominant for one gene, in order to determine the homozygous individuals, by using a test cross to a recessive individual. The progenies of the homozygotes will not show segregation. The progenies of the heterozygotes are expected to segregate into one-half dominants (**Aa**) and one-half recessives (**aa**). The only error will arise from the failure of the progenies of some heterozygotes to contain at least one recessive. Let it further be decided that such a misleading result, i.e. failure of segregation in the progeny of a heterozygote, must not occur with a frequency of more than 1 per cent on the average.

Now in the progeny of a heterozygote each individual has a chance of $\frac{1}{2}$ of being a dominant. Then a family of n individuals will all be dominant in $(\frac{1}{2})^n$ of cases. This is the misleading result and must not occur in more than 1 per cent of cases. Then the minimum value of n is given by the solution of the equation

$$(\tfrac{1}{2})^n = \tfrac{1}{100}$$

Taking logarithms this becomes $n \log (\frac{1}{2}) = \log (\frac{1}{100})$

$$\text{i.e.} \qquad -0{\cdot}3010n = -2{\cdot}000$$

$$\text{or} \qquad n = \frac{2}{0{\cdot}3010} = 6{\cdot}6.$$

The minimum size of the progeny must be 7.

If we had been dealing with a case of two factors, i.e. where the individuals for testing could have been heterozygous for as many as two genes, a family of n individuals, n having been chosen for a certain probability of failing to show segregation of one factor heterozygous in the parent, would be twice as likely to fail to show segregation for two factors. Then in such cases we must increase the stringency of our

test in order to allow for this fact. In the above example if we had been concerned with two factors in the test backcrosses we should have equated $(\frac{1}{2})^n$ to $\frac{1}{200}$, instead of $\frac{1}{100}$, and so obtained 7·6, or 8, as the minimum size of n in the test.

Another and very similar problem is answered in the same way. Suppose, for example, we have an F_2 family segregating for two genes and we want to breed from a homozygous doubly dominant individual. Then how many phenotypically **AB** individuals must be used in order to include at least one **AABB**, the maximum frequency of failure to be $\frac{1}{500}$?

Now out of every nine individuals of the pheno type **AB** in such an F_2 we expect one to be **AABB**. Then the chance of an individual being heterozygous for at least one factor is $\frac{8}{9}$. The chance of *all* of n being so heterozygous is $(\frac{8}{9})^n$. The maximum allowable failure is $\frac{1}{500}$.

Then
$$(\tfrac{8}{9})^n = \tfrac{1}{500}$$
$$n \log (\tfrac{8}{9}) = \log (\tfrac{1}{500})$$
$$-0{\cdot}0512n = -2{\cdot}6990$$
and
$$n = \frac{2{\cdot}6990}{0{\cdot}0512} = 52{\cdot}7$$

At least fifty-three phenotypically **AB** individuals should be used.

At the end of the book will be found a table (Table III) giving a series of such minimum numbers of progeny. The leftmost column shows the fraction of the individuals, which are expected to be of the distinctive type (e.g. it would be $\frac{1}{2}$ in the first and $\frac{1}{9}$ in the second of the above examples) and along the top is the precision of the test. Thus the last example could be found in the table by looking along the row corresponding to $\frac{1}{9}$ until the column of 0·998 precision, i.e. 0·002 (i.e. $\frac{1}{500}$) error, is reached. The value in the table is then the minimum value of n, the size of the progeny. As a further example of the use of this table, consider the question of the

minimum size of family required to obtain at least one individual of a type which is expected to comprise $\frac{1}{4}$ of the total, the maximum error to be $\frac{1}{50}$. Then we *look along row* $\frac{1}{4}$ *and down column* 0·980 (i.e. $\frac{49}{50}$ precision) and find the value of 13·6 for n. At least fourteen individuals should be grown.

10. DISTINGUISHING TWO SEGREGATIONS

A problem superficially related to the foregoing but treated differently is that of determining the number of individuals necessary in order to decide between two different types of segregation, and consequently between the two hypotheses on which the segregation expectations are based. There must be some minimum probability laid down for this decision too.

As an example, suppose we wished to decide whether one or both of two complementary factors was segregating in an F_2. If only one were segregating, the dominant allelomorph of the other being present in all individuals, a 3 : 1 ratio would be found. If both were segregating, a 9 : 7 would be obtained.

Let n be the size of the F_2 necessary for our purpose. There will be some number (r) of recessives, which if occurring in a family of size n will leave both hypotheses equally likely. If more than r recessives occur, then the 9 : 7 ratio is more likely, and if less than r recessives occur, the 3 : 1 is favoured. Hence, to solve the problem we make the family sufficiently large to ensure that r recessives, if found, will show a deviation from expectation on either hypothesis of a size that could occur only with that probability chosen as the maximum for misclassification. If the number of recessives is other than r, as indeed it must usually be, then one or other hypothesis is less likely than the maximum misclassification allowed, and so may be judged to be incorrect.

The actual calculation may be done in either of

two ways, one using the standard error test of significance and the other the χ^2 test. Let us consider the standard error method first.

The standard error of the number of recessives expected with a 3 : 1 ratio is $\sqrt{\frac{3n}{16}}$ and with an expectation of 9 : 7 is $\sqrt{\frac{63n}{256}}$.

Now if we take 0·025 as the maximum allowable misclassification, we actually utilize the deviate corresponding to 0·05, because deviation in but one of the two possible directions is misleading. We find, from Table I, that the deviation of r from the expected number of recessives must not be less than 1·959964 times the standard error.

Then, for the 9 : 7 expectation,

$$\tfrac{7}{16}n - r = 1{\cdot}959964\sqrt{\frac{63n}{256}}$$

and for the 3 : 1 expectation

$$r - \tfrac{1}{4}n = 1{\cdot}959964\sqrt{\frac{3n}{16}}$$

Then by addition

$$n\left(\tfrac{7}{16} - \tfrac{1}{4}\right) = 1{\cdot}959964\sqrt{n}\left(\sqrt{\tfrac{63}{256}} + \sqrt{\tfrac{3}{16}}\right)$$

and $$\sqrt{n} = \frac{16}{3}\left[1{\cdot}959964\left(\frac{7{\cdot}937254}{16} + \frac{1{\cdot}732051}{4}\right)\right]$$

$$= 9{\cdot}711919$$

$$n = 94{\cdot}32.$$

The method of χ^2 should give the same answer if applied correctly. For this approach it is necessary to note that we have two expected segregations, $l_1 : 1$ and $l_2 : 1$. Then the observed segregation which will give equal χ^2s on both hypotheses is $\sqrt{l_1 l_2} : 1$.

In our case $l_1 = 3$ and $l_2 = \frac{9}{7}$, and so the ambiguous

segregation containing r recessives is, $\sqrt{\frac{27}{7}}:1$ i.e. $1{\cdot}9640:1$ and $r = \frac{n}{2{\cdot}9640}$.

Now taking the ambiguous segregation of

$$\frac{1{\cdot}9640n}{2{\cdot}9640} : \frac{n}{2{\cdot}9640}$$

and calculating χ^2 on the hypothesis of 3 : 1 we **find**, using the formula given in Section 8,

$$\chi^2 = \frac{\left[\frac{1{\cdot}9640n}{2{\cdot}9640} - \frac{3n}{2{\cdot}9640}\right]^2}{3n} = 3{\cdot}841$$

for the 0·025 level of deviation probability. We take $\chi^2 = 3{\cdot}841$ for a probability of 0·025 as deviations in but one of the two possible directions are misleading.

Then $$\frac{n^2}{(2{\cdot}9640)^2} \times \frac{1}{3n}[1{\cdot}9640 - 3]^2 = 3{\cdot}841$$

or $$n = \frac{3{\cdot}841 \times 3 \times (2{\cdot}9640)^2}{(1{\cdot}9640 - 3)^2}$$

$$= \frac{101{\cdot}230}{1{\cdot}073} = 94{\cdot}31$$

This answer differs by about 0·01 per cent from the previous one—an error within the limits allowed by calculation.

Thus we must grow ninety-five plants in order to distinguish between the two hypotheses with a minimum certainty of 0·025.

CHAPTER IV

THE DETECTION OF LINKAGE

11. ANALYSIS OF χ^2 BY ORTHOGONAL FUNCTIONS

HAVING dealt with the single factor ratios, attention may now be turned to the detection of linkage. It will be assumed for the present that no complications are introduced by aberrant single gene segregations. The cases where the single factors are not giving segregations in strict agreement with Mendelian expectation will be dealt with later (Chap. VIII).

The method of analysis by χ^2 is the most profitable approach to the detection of linkage. The procedure for the calculation of χ^2 is essentially similar to that appropriate to the single factor ratios.

Let us consider a backcross involving two factors. One parent is **Aa Bb**, i.e. doubly heterozygous, and the other is the double recessive **aa bb**. If the two factors are each segregating in accordance with the Mendelian expectation of 1 : 1 and provided that there is no linkage, we expect four classes of offspring, **AaBb, Aabb, aaBb, aabb,** in equal numbers. Where m_1 is the expectation for the first class, m_2 for the second and so on,

$$m_1 = m_2 = m_3 = m_4 = \tfrac{1}{4}$$

Further, let $a_1 \cdots a_4$ be the observed frequencies of the four classes, the total being represented by n.

In the first place it is possible to calculate a χ^2 for the joint deviation of all the observed frequencies

from their expectations, by the use of the formula $S\left(\frac{a^2}{mn}\right) - n$. This χ^2 has three degrees of freedom, as there are four classes of which three may be filled arbitrarily. It must include two components which correspond to the deviations of each of the two single factor ratios from their expectations, and one for the joint segregation from its expectation of no linkage. Our task is clearly to separate these components in such a way as to allow of separately testing the three possible sources of discrepancy. The three degrees of freedom can be conveniently subdivided into

1 for the deviation of the **Aa** segregation from 1 : 1
1 ,, ,, ,, ,, ,, **Bb** ,, ,, 1 : 1

and 1 detecting association of the two factors in segregation, our expectation or null hypothesis being, of course, that they are independent.

The two χ^2s corresponding to the first two degrees of freedom are calculated by the methods given in the last chapter. In this case

$$\chi^2_A = \frac{(a_1 + a_2 - a_3 - a_4)^2}{n}$$

and

$$\chi^2_B = \frac{(a_1 - a_2 + a_3 - a_4)^2}{n}$$

It is easily seen that these reduce to the corresponding formulae of Section 8.

The formula for the calculation of that χ^2 value corresponding to the third, or 'linkage', degree of freedom follows from the two already employed by application of the principle of orthogonality and is

$$\chi^2_L = \frac{(a_1 - a_2 - a_3 + a_4)^2}{n}$$

Any other formula would yield a χ^2 which would be based on a comparison itself not independent of the two comparisons already used. This would clearly

defeat our object of testing the three sources of deviation separately. The principle of orthogonality and the arrangement of formula based on independent comparisons is dealt with more thoroughly in the last section of this chapter. Having obtained these three separate χ^2 values each with one degree of freedom it is possible to test the three sources of discrepancy individually.

Ex. 5. As an example of this type of analysis we may take the data of Philp (1934) on the joint segregation of the two factors **p** and **t** in the poppy. In a backcross progeny $\left(\frac{\mathbf{pt}}{\mathbf{PT}} \times \frac{\mathbf{pt}}{\mathbf{pt}}\right)$ he observed the following classes and frequencies.

TABLE 5

	PpTt	Pptt	ppTt	pptt	Total
Observed . .	191	37	36	203	467
Expected (with no linkage) .	116·75	116·75	116·75	116·75	467

A χ^2 for three degrees of freedom as calculated from these four classes as they stand has the value of 221·266. This is clear indication of strong deviation from the expectation of equal classes. To what is the deviation due ? The next step is to subdivide χ^2 into its three components.

First take the deviation of **P**,**p** segregation from the 1 : 1. This component of χ^2 is found by adding the first and second classes together and the third and fourth classes together, taking the difference and basing the calculation on this. The formula in the previous notation is

$$\chi^2{}_P = \frac{(a_1 + a_2 - a_3 - a_4)^2}{n}.$$

χ^2 is then found to be 0·259.

Similarly χ^2 for the deviation of the **T**, **t** segregation from 1 : 1 is found to be 0·362.

The third component, that detecting linkage, is

based on the difference between the sums of the first and fourth classes and the second and third classes. The formula is, as before,

$$\chi^2{}_L = \frac{(a_1 - a_2 - a_3 + a_4)^2}{n} = \frac{(191 - 37 - 36 + 203)^2}{n}$$

and this component proves to be 220·645.

The three components and their probabilities may now be tabulated as in Table 6.

TABLE 6

		D.F.	Probability
Segregation for **P,p**	0·259	1	0·7 – 0·5
Segregation for **T,t**	0·362	1	0·7 – 0·5
Joint segregation	220·645	1	extremely small
Total	221·266	3	

The total of the three components agrees with the compound χ^2 previously calculated by a different method, so demonstrating that the working is correct.

It is clear from this partition of χ^2 that the two single factor ratios are individually good, but that there is very strong evidence for the belief that the factors are not segregating independently of one another. The two dominant allelomorphs and the two recessive allelomorphs are associated too often, or, in other words, there is very strong evidence for the existence of linkage in the coupling phase.

This same method of analysis may be applied to data obtained from inbreeding doubly heterozygous individuals. In this case the four classes are expected to occur in the ratio 9 : 3 : 3 : 1. The formulae for the three components in this type of family are somewhat different from those used in the case of the backcross. The two components corresponding to the single factor ratios are calculated from the formulae for the single factor ratios with expectation 3 : 1 (cf. Section 8). The third component then

follows from orthogonality. The three formulae are :

$$\chi^2{}_A = \frac{(a_1 + a_2 - 3a_3 - 3a_4)^2}{3n} \quad \chi^2{}_B = \frac{(a_1 - 3a_2 + a_3 - 3a_4)^2}{3n}$$

$$\chi^2{}_L = \frac{(a_1 - 3a_2 - 3a_3 + 9a_4)^2}{9n}$$

Where a number of families of the same type are available it is possible to subdivide χ^2 not only into the three component parts discussed above but also into portions depending on deviation of the totals and on heterogeneity respectively as in the case of the single factor ratios previously considered. This will be made clear by an example.

Ex. 6. The joint segregation of the two factors **A,a** and **B,b** in *Pharbitis*, Morning Glory, has been studied by Imai (1931). He records the segregations in three families as shown in Table 7.

TABLE 7

	AB	Ab	aB	ab	Total
Family 1 . .	47	8	11	9	75
,, 2 . .	75	14	14	11	114
,, 3 . .	65	13	12	11	101
Total . .	187	35	37	31	290

First consider the totals of the combined families as given in the bottom row of the table. The three components of χ^2 are calculated from the formulae shown above. The following analysis is then obtained :

	χ^2	D.F.	Probability
Segregation for **A,a** .	0·372	1	0·5 – 0·3
,, ,, **B,b** .	0·777	1	0·5 – 0·3
Joint Segregation .	23·946	1	very small
Total.	25·095	3	

It is thus quite clear that again the single factor ratios account for very little of the total χ^2 but that

there is a large component corresponding to linkage. There is undoubtedly evidence for linkage of the two segregating factors.

This process of analysis may be carried out for each family separately. In each case there will be three χ^2s each corresponding to one degree of freedom. This is carried out and tabulated in Table 8.

TABLE 8

Family	χ^2 Factor pair A,a	χ^2 Factor pair B,b	χ^2 Linkage
1	0·111	0·218	7·468
2	0·573	0·573	7·895
3	0·267	0·083	8·714
Total	0·951	0·874	24·077

The bottom row of the same gives the sums, over all three families, for each component. Each sum has three degrees of freedom, one being contributed by each family. The analysis into deviation and heterogeneity portions is now carried out as in the examples of Chapter II. The deviation portion has already been obtained in the previous table. The difference between this and the corresponding total from Table 8 is the heterogeneity χ^2. In each case the deviation χ^2 will have one degree of freedom and the heterogeneity χ^2 will have $3 - 1$, i.e. two degrees of freedom. This partition is shown in Table 9.

TABLE 9

	A,a	B,b	Linkage	D.f.
Deviation	0·372	0·777	23·946	1
Heterogeneity	0·579	0·097	0·131	2
Total	0·951	0·874	24·077	3

The heterogeneity χ^2 are none of them significant. We can now add the further statement that the families are homogeneous for each component. They agree in showing good single factor ratios and they

also agree in showing linkage of the two factors. It may be noted that in the case of the linkage component the heterogeneity χ^2 will be somewhat too low as has earlier been shown to be the case with single factor ratios when a significant deviation is recorded. The difference between the value found for this linkage heterogeneity χ^2 and the true value will, however, be small and since the value found above is very small we can assume that the true value will not reach the level of significance. The Brandt and Snedecor technique cannot be applied to finding the true value, as χ^2 is calculated from four classes weighted in different manners.

Thus the use of the χ^2 test of goodness of fit allows of analysis of the data which not only detects irregularities but also shows precisely where the irregularities occur. The presence of linkage is often obvious as in the data worked in Ex. 5, but this is not always the case. Before its presence is assumed, a sensitive statistical test, as is provided by χ^2, should be applied. In the single families of Ex. 6 the total χ^2 which corresponds to three degrees of freedom shows a barely significant deviation from expectation (e.g. Family 1 $\chi^2 = 7{\cdot}798$ D.F. $= 3$ Probability $0{\cdot}05 - 0{\cdot}02$). The advantage of the analysis in such a case lies in its showing that two of the three possible sources of deviation, the single factor ratios, contribute very little to χ^2 whereas the third, linkage, contributes much. The sensitivity of the test for linkage is thus very greatly increased.

12. ORTHOGONALITY

The successful analysis of χ^2 into its components depends on the choice of functions which give independent comparisons. That this must be so, is clear when it is remembered that χ^2 is analysed in order to locate the precise place in which the results fail to conform to expectation. It would be inefficient, for the detection of linkage, to calculate some χ^2

value which is not independent of single factor segregation on the hypothesis of no linkage.

In the two examples worked above the reason given for the choice of the functions used in calculating the χ^2 components was that such functions were orthogonal. The following discussion of orthogonality will help to make this choice clear.

Consider a segregation into four classes of expectation m_1 to m_4 and with observed frequencies of a_1 - - - a_4. Various linear functions of the observed frequencies may be taken, the general form being

$$x = k_1a_1 + k_2a_2 + k_3a_3 + k_4a_4.$$

Where V_x is the random sampling variance of the linear function x, it can be shown that $\frac{x^2}{V_x}$ is distributed as a χ^2 for one degree of freedom. If the coefficients of a_1, &c., are chosen correctly the resulting χ^2 will detect deviations from some specific expectation of the class frequencies. In the case where $m_1 = 3m_2 = 3m_3 = 9m_4$, i.e. in an F_2 family segregating for two factors, and the choice of $k_1 = k_2 = 1$ and $k_3 = k_4 = -3$ is made, the resulting χ^2 will detect deviations of one single factor segregation from the expected 3 : 1. When the expected value of x, i.e. $k_1m_1 + k_2m_2 + k_3m_3 + k_4m_4$, is 0 the value of V_x is obtained from the formula $\frac{1}{n} V_x = S(mk^2)$ where n is the number of individuals in the family. If the expected value of x is not 0 this formula for V_x ceases to hold. So the coefficients should be chosen to make the expectation of x zero. In the case mentioned above

$k_1m_1 = \frac{9}{16}$, $k_2m_2 = \frac{3}{16}$, $k_3m_3 = -\frac{9}{16}$, $k_4m_4 = -\frac{3}{16}$
and so $S(km) = 0$.

$$\text{Then } \frac{1}{n}Vx = S(mk^2) = \tfrac{1}{16}(9 + 27 + 3 + 9) = 3$$

$$\therefore\ Vx = 3n \text{ and } \chi^2 = \frac{x^2}{3n}$$

Any number of such functions may be chosen, but they will not all be orthogonal. Orthogonality is tested by calculating the quantity $S(mkk')$. This should be zero. If it is not found to be so the two functions are not based on independent comparisons.

In addition to the function taken above, let us take a second one, x', where $k'_1 = k'_3 = 1$ and $k'_2 = k'_4 = -3$. Then $S(mk') = \frac{1}{16}(9 - 9 + 3 - 3) = 0$

$$Vx = n(Smk'^2) = \frac{n}{16}(9 + 27 + 3 + 9) = 3n$$

$\frac{x'^2}{3n}$ will then be distributed as χ^2.

Furthermore,

$$S(mkk') = \frac{1}{16}(9 - 9 - 9 + 9) = 0$$

and so this and the previous function are orthogonal. In point of fact they are based on the single factor ratios of the two genes in the segregation. Thus we have taken two functions each giving a χ^2 of one degree of freedom. The third component still remains to be determined. It will be of the nature of what is termed, in factorial experimentation, an 'interaction', as it will detect association of the two primary factors in segregation. The coefficients of the observed class frequencies must be chosen to make this third functional orthogonal to the other two. They are easily found by a multiplication process.

$$k''_1 = k_1k'_1 = 1 \times 1 = 1$$
$$k''_2 = k_2k'_2 = 1 \times -3 = -3$$
$$k''_3 = k_3k'_3 = -3 \times 1 = -3$$
$$k''_4 = k_4k'_4 = -3 \times -3 = 9$$

Then $x'' = a_1 - 3a_2 - 3a_3 + 9a_4$, with $S(mk'') = 0$

and $Vx'' = nS(mk''^2) = \frac{n}{16}(9 + 27 + 27 + 81) = 9n$

χ^2 is given by $\frac{x''^2}{9n}$.

The orthogonality is tested as before and we find

$$S(mkk'') = \tfrac{1}{16}(9 - 9 + 27 - 27) = 0$$
$$S(mk'k'') = \tfrac{1}{16}(9 + 27 - 9 - 27) = 0$$

Thus the analysis of χ^2 in an F_2 family is conducted by calculating the quantities

$$\chi^2 = \frac{1}{3n}(a_1 + a_2 - 3a_3 - 3a_4)^2$$

$$\chi^2 = \frac{1}{3n}(a_1 - 3a_2 + a_3 - 3a_4)^2$$

$$\chi^2 = \frac{1}{9n}(a_1 - 3a_2 - 3a_3 + 9a_4)^2$$

The functions chosen for the calculation of the first two χ^2 are the same as those developed in the previous chapter from the detection of deviations from single factor ratios. The third function, detecting linkage, then follows from the above considerations. Other sets of three orthogonal functions could be chosen but would not have the same meaning for the analysis as do those adopted above.

The three functions chosen for the analysis of the backcross data in Ex. 5 can be derived in the same way as the functions for the F_2. Similarly it can be shown that in the case of two factors one of which is a member of two segregating duplicate genes, giving four classes with the expected ratios $45 : 15 : 3 : 1$, the three components of χ^2 are calculated from the formulae

$$\chi^2_A = \frac{1}{3n}(a_1 - 3a_2 + a_3 - 3a_4)^2$$

$$\chi^2_B = \frac{1}{15n}(a_1 + a_2 - 15a_3 - 15a_4)^2$$

$$\chi^2_L = \frac{1}{45n}(a_1 - 3a_2 - 15a_3 + 45a_4)^2$$

The derivation of these functions, and the demonstration of their orthogonality, is left as an exercise.

A more complicated example is that of a backcross for three factors. Eight equal classes are expected, provided the genes are unlinked and showing good individual segregation ratios. There are seven degrees of freedom and so seven χ^2s can be calculated. One set of seven orthogonal χ^2s, and the most useful set for genetical analysis, is obtained by giving the classes coefficients as shown in Table 10. The variance of each function can be shown to be n. The χ^2s are then calculated by squaring the quantities derived from the Table 10 and dividing by n in each case.

The quantities in the table are orthogonal and are easily derived by the multiplication method previously employed. The first three (1, 2, and 3) are those which detect deviations from 1 : 1 in the three single factor ratios. They are made up by giving all the classes, dominant for the factor under consideration, a coefficient of 1 and all classes recessive for the factor a coefficient of − 1. The interaction or linkage function for factors **Aa** and **Bb** (4) is then obtained by multiplying the coefficients of the single gene ratio functions of these two together (1 and 2). Similar linkage functions are obtained for **A,a** and **C,c**, and **B,b** and **C,c**. The seventh and last degree of freedom corresponds to a function which has no simple genetical meaning but which is necessary to complete the analysis. The coefficients of this function are obtained by multiplying those of the first (**A,a**) single gene ratio function (1) by those of the function corresponding to linkage between **B,b** and **C,c** (6). It may be also obtained from multiplication of the second (**B,b**) function (2) by the linkage function of **A,a** and **C,c** (5) and finally by the corresponding multiplication of the third and fourth functions (3 and 4) of the table. This multiplication method of obtaining the orthogonal functions corresponding to linkage degrees of freedom may be employed in any case that may arise.

TABLE 10

BACKCROSS FOR THREE FACTORS

Coefficients of the Functions for Calculation of χ^2

Function Genetical Class	1	2	3	4	5	6	7
AaBbCc	1	1	1	1	1	1	1
AaBbcc	1	1	−1	1	−1	−1	−1
AabbCc	1	−1	1	−1	1	−1	−1
Aabbcc	1	−1	−1	−1	−1	1	1
aaBbCc	−1	1	1	−1	−1	1	−1
aaBbcc	−1	1	−1	−1	1	−1	1
aabbCc	−1	−1	1	1	−1	−1	1
aabbcc	−1	−1	−1	1	1	1	−1
Total .	0	0	0	0	0	0	0

Function 1 gives χ^2 for the segregation of the single factor **A,a**.
,, 2 ,, ,, ,, ,, ,, ,, **B,b**.
,, 3 ,, ,, ,, ,, ,, ,, **C,c**.
,, 4 ,, ,, linkage between **A,a** and **B,b**.
,, 5 ,, ,, ,, ,, **A,a** ,, **C,c**.
,, 6 ,, ,, ,, ,, **B,b** ,, **C,c**.
,, 7 completes the analysis.

CHAPTER V

THE ESTIMATION OF LINKAGE

13. CRITERIA OF ESTIMATION

HAVING detected the presence of linkage in a segregation for two or more genes the next step is, naturally, to obtain some measure of the intensity of the linkage. It should be noted at this point that though detection of linkage involves no hypothesis as to the nature of linkage, being merely the demonstration that the hypothesis of free segregation is not true, the measurement of the intensity of linkage involves the calculation of a statistic relevant to some hypothesis of its nature. For example, the measure of linkage is a different one if the chromosome theory is accepted from that measure used if one adopts the gametic reduplication hypothesis. The chromosome theory, which is generally accepted, leads to a measure of the intensity of linkage based on the frequency of breakage and rejoining of the homologous chromosomes between the loci concerned. This is estimated as the proportion of recombination chromosomes. In the case of diploid organisms, this is the same as estimating the frequency of recombination gametes.

Having thus decided on the quantity, or parameter, as it is termed, which is to be estimated, the next decision to be taken is as to the method of estimation. Two criteria must be satisfied in the case of linkage estimation. The first, that of consistency, concerns the statistic itself. Care must be taken that this is

really an estimate of the parameter concerned and not of something different. This appears, at first sight, to be obvious but it must be remembered that different types of data lead to estimates of different things. The backcross allows of direct estimation of the recombination fraction, whereas F_2 data can at best only give estimates of either the square of the recombination fraction, or of the square of one minus this fraction. Where the recombination fraction differs on the male and female sides, the product of the two fractions, or of one minus each fraction, is estimated. To overlook this possibility would be most misleading.

The second criterion concerns the precision of the estimate. We must take care to obtain the most precise estimate possible, in the sense that the estimate should have the smallest variance, or standard error, that the data can give. This is the criterion of efficiency. The reasons for the application of this criterion will be made clear later. In some cases an inefficient estimate may be employed if the efficient estimate is difficult to obtain, but this is a course which should not generally be followed. A third criterion, that of sufficiency (see Fisher, 1949*a*) is involved in some cases of estimation but can be neglected in the case of linkage. Furthermore, the method given below for linkage estimation will lead to a sufficient estimate, if one exists.

The satisfaction of the criterion of consistency is a matter of choosing the right quantity to estimate. This will be illustrated fully by the examples worked below. Satisfaction of the criterion of efficiency is one which can only be considered mathematically. It will be sufficient to say here that the method given below has been shown always to lead, in the theory of large samples, to an efficient statistic, i.e. an estimate having the smallest standard error of which the data will allow.

This method is that of Maximum Likelihood.

The principle of the method is easily grasped. Let p be the recombination fraction, $m_1 \ldots m_t$, the expected proportions of individuals in the segregate classes $1 \ldots t$, and $a_1 \ldots a_t$ the numbers of individuals observed in these classes. The expected proportions, denoted by m, are known in terms of p, the quantity which is to be estimated.

The likelihood of obtaining the observed family is given by a term of the expansion of

$$(m_1 + m_2 \ldots + m_t)^n$$

where n is the total individuals in the family (see Fisher, 1921). The relevant term is,

$$\frac{n!}{a_1!a_2! \ldots a_t!} (m_1)^{a_1} (m_2)^{a_2} \ldots (m_t)^{a_t}$$

The method of maximum likelihood depends on the maximization of this expression, with respect to p. It is difficult, however, to differentiate such an expression and resort is made to a device for this purpose. The expression and its logarithm will both be maxima at the same value of p. Hence we may find the requisite recombination fraction by maximizing the logarithm of the likelihood expression with respect to p.

The logarithm of the likelihood expression, denoted by L, is

$$L = C + a_1 \log m_1 + a_2 \log m_2 + \ldots a_t \log m_t$$

where C is a constant depending on the coefficient of the likelihood term. This will vanish on maximization by differentiation and so may be neglected.

Differentiating and equating to zero leads to the equation of estimation

$$\frac{dL}{dp} = a_1 \frac{d \log m_1}{dp} + a_2 \frac{d \log m_2}{dp} \ldots + a_t \frac{d \log m_t}{dp} = 0$$

One of the solutions of this equation will be the desired value of p. There is never any doubt as to which root is required since all the others lead to impossible values for the recombination fraction.

14. THE CALCULATION OF MAXIMUM LIKELIHOOD

• *Ex.* 7. Let us consider the estimation of the recombination fraction p in the case of the factors **p** and **t** in the poppy (Philp's data). It has been shown in the previous chapter (Ex. 5) that the data give evidence of linkage of these factors. The cross was one of the double heterozygote in the coupling phase to the double recessive $\left(\frac{\mathbf{pt}}{\mathbf{PT}} \times \frac{\mathbf{pt}}{\mathbf{pt}}\right)$. All the gametes from the recessive parent will be **pt** and so do not enter into our consideration of the problem. The gametes from the heterozygous parent will be of four kinds, two of which are old, or original, combinations and the other two new, or re-, combinations. In this way the expected frequencies shown in Table 11 are arrived at. The corresponding observed numbers are also shown.

TABLE 11

	PT	Pt	pT	pt	Total
Observed .	191	37	36	203	467
Expected .	$\frac{n}{2}(1-p)$	$\frac{n}{2}p$	$\frac{n}{2}p$	$\frac{n}{2}(1-p)$	n

The logarithm likelihood expression is thus

$$L = 191 \log (\tfrac{1}{2} - \tfrac{1}{2}p) + 37 \log (\tfrac{1}{2}p) + 36 \log (\tfrac{1}{2}p) + 203 \log (\tfrac{1}{2} - \tfrac{1}{2}p)$$

and maximizing by differentiation and equating to zero the equation of estimation becomes :

$$\frac{dL}{dp} = -\frac{191}{1-p} + \frac{37}{p} + \frac{36}{p} - \frac{203}{1-p} = 0$$

The solution is $p = \frac{73}{467} = 0{\cdot}1563$ or 15·63 per cent.

It will be noticed that in the case of the backcross this method of estimation leads to the formula which is in universal use for data of this kind, viz.

$$p = \frac{a_2 + a_3}{n} \left[\text{or } \frac{a_1 + a_4}{n} \text{ in the case of repulsion}\right]$$

Having obtained our estimate of p, we are next concerned with its standard error (s_p). Where V_p is the variance, i.e. the standard error squared, of p, it can be shown that

$$-\frac{1}{V_p} = S\left(mn\frac{d^2 \log m}{dp^2}\right)$$

This is easily calculated for the present example. We already have $a\frac{d \log m}{dp}$. We must differentiate for a second time and then substitute the expected for the observed value, i.e. nm for a. This gives

$$-\frac{1}{V_p} = -\frac{n}{2}\left(\frac{1}{1-p} + \frac{1}{p} + \frac{1}{p} + \frac{1}{1-p}\right)$$

$$= -\frac{n}{p(1-p)} = -\frac{467}{p(1-p)}$$

Inserting the estimated value of p we obtain $V_p = 0{\cdot}0002824$ and $s_p = \sqrt{V_p} = 0{\cdot}0168$ or $1{\cdot}68$ per cent.

The general formula for s_p in the case of a back-cross is $\sqrt{\frac{p(1-p)}{n}}$ which is the formula in universal use. It is of some interest that, in this simple case, the formula given by the method of maximum likelihood is that previously arrived at by the application of simpler statistical considerations.

Ex. 8. As a further example of estimation, we may consider the data of Imai concerning the genes **A,a** and **B,b** in *Pharbitis*. These were shown earlier (Ex. 6) to show strong evidence of linkage in the coupling phase. These data are from F_2 families. The first thing is to ascertain the expectations of the four observed classes in terms of p. The gametic series of expectations will be the same as the series expected in the case of the backcross. But we have

no reason to assume that the recombination fraction is the same in both male and female gametogenesis. We may represent the former by p_1 and the latter by p_2. The gametic series of expectations are then :

TABLE 12

	AB	Ab	aB	ab
♂ . . .	$\frac{1}{2}(1 - p_1)$	$\frac{1}{2}p_1$	$\frac{1}{2}p_1$	$\frac{1}{2}(1 - p_1)$
♀ . . .	$\frac{1}{2}(1 - p_2)$	$\frac{1}{2}p_2$	$\frac{1}{2}p_2$	$\frac{1}{2}(1 - p_2)$

From this table it is possible to build up the expectations of the four phenotypically distinct F_2 classes. The double recessive class can only result from mating of doubly recessive male and female gametes and so will be expected in $\frac{1}{4}(1 - p_1)(1 - p_2)$ of cases. The total incidence of **a** plants is $\frac{1}{4}$ and so the singly recessive class, **aB**, will occur in $\frac{1}{4}[1 - (1 - p_1)(1 - p_2)]$. The other singly dominant class will have equal expectation and the doubly dominant class, **AB**, must be $\frac{1}{4}[2 + (1 - p_1)(1 - p_2)]$. Thus all the expectations are dependent on the quantity $(1 - p_1)(1 - p_2)$. This is the parameter which can be estimated. If we care to assume that $p_1 = p_2$ it is possible to obtain an estimate of p, but only if this assumption is made.

Let us write P for $(1 - p_1)(1 - p_2)$. Then the expectations of the four classes are as shown in Table 13. The observed frequencies of the four classes are also shown in that table.

TABLE 13

Class	AB	Ab	aB	ab
Expectation .	$\frac{n}{4}(2 + P)$	$\frac{n}{4}(1 - P)$	$\frac{n}{4}(1 - P)$	$\frac{n}{4}P$
Observed . .	187	35	37	31

The logarithm likelihood expression is then :

$$L = 187 \log (\tfrac{1}{2} + \tfrac{1}{4}P) + 35 \log (\tfrac{1}{4} - \tfrac{1}{4}P) + 37 \log (\tfrac{1}{4} - \tfrac{1}{4}P) + 31 \log \tfrac{1}{4}P$$

Maximization leads to the equation :

$$\frac{dL}{dP} = \frac{187}{2+P} - \frac{35}{1-P} - \frac{37}{1-P} + \frac{31}{P} = 0$$

which reduces to the quadratic

$$62 + 12P - 290P^2 = 0$$

giving $P = 0{\cdot}4835$.

The variance of P is to be obtained by the method given previously. Redifferentiating and substituting nm for a

$$-\frac{1}{V_P} = -\frac{n}{4}\left(\frac{1}{2+P} + \frac{1}{1-P} + \frac{1}{1-P} + \frac{1}{P}\right)$$

$$V_P = \frac{2P(1-P)(2+P)}{n(1+2P)}$$

Hence $V_P = 0{\cdot}002174$ and $s_p = 0{\cdot}04663$.

If we now care to assume that $p_1 = p_2$ we have

$$(1-p) = \sqrt{P} = 0{\cdot}6953$$
$$p = 0{\cdot}3047 \text{ or } 30{\cdot}47 \text{ per cent.}$$

The variance of p is then found from the variance of P. It can be shown that

$$\frac{1}{V_p} = \frac{1}{V_P}\left(\frac{dP}{dp}\right)^2$$

and since $P = (1-p)^2$, $\left(\frac{dP}{dp}\right)^2 = 4P$.

Hence $$V_p = \frac{V_P}{4P} = 0{\cdot}001124$$

and $s_p = \sqrt{V_p} = 0{\cdot}03353$ or $3{\cdot}35$ per cent.

It will be noticed that linkage could be detected by the calculation of p and its standard error and then testing the significance of the deviation of p from the freedom value of 0·5. This method, if correctly applied, should give the same result as the use of the χ^2 method. It has, however, the disadvantage of not leading to such a fine analysis of

the situation, in respect of heterogeneity tests, &c., as does the χ^2 test. The latter is to be preferred for detection.

15. THE EFFICIENCY OF STATISTICS

It may be wondered at that the method of estimation bears no relation to the calculation of χ^2 which is so good for detection. It would superficially seem reasonable to employ the linear function x from which χ^2 is calculated as a method of estimating the recombination fraction. Let us consider the estimation of P from an F_2 family using this method. Where the four observed classes are $a_1 \ . \ . \ . \ a_4$ the linkage function for the calculation of χ^2 is

$$x = a_1 - 3a_2 - 3a_3 + 9a_4$$

The expected value of this function in terms of P is

$$x = \frac{n}{4}(16P - 4)$$

Then P may be estimated from the equation

$$n(4P - 1) = a_1 - 3a_2 - 3a_3 + 9a_4$$

or $$P = \frac{1}{4n}(2a_1 - 2a_2 - 2a_3 + 10a_4)$$

We are next concerned with the estimation of the standard error of P arrived at by this method. The sampling variance of such a statistic is obtained from the formula

$$nV_P = S(mk^2) - P^2 \text{ (Fisher 1949}a\text{)}$$

which is related to the formula used for obtaining the sampling variance when calculating χ^2. The chief difference is due to the fact that in this case $S(mk)$ does not equal 0.

Here

$$m_1 = \tfrac{1}{4}(2 + P) \quad m_2 = m_3 = \tfrac{1}{4}(1 - P) \quad m_4 = \tfrac{1}{4}P$$

and $$k_1 = \tfrac{1}{2} \quad k_2 = k_3 = -\tfrac{1}{2} \quad k_4 = 2\tfrac{1}{2}$$

$$\therefore \ 4nV_P = \tfrac{1}{4}(4 + 24P) - 4P^2$$

and so $$V_P = \frac{1 + 6P - 4P^2}{4n}$$

This is not the same as the variance of the maximum likelihood statistic. Now the smaller the variance the more precise the estimate and so the more *efficient* the statistic. The maximum likelihood statistic has the smallest possible variance and so is always 100 per cent efficient. The efficiency of any other statistic may conveniently be expressed as the ratio of the variance of the maximum likelihood statistic to the variance of the statistic in question. In the case of Imai's data the efficiency is then

$$\frac{8P(1 - P)(2 + P)}{(1 + 6P - 4P^2)(1 + 2P)}$$

P was found to be 0·4835 and so the efficiency is 0·8505 or 85·05 per cent.

Where the value of P is $\frac{1}{4}$, i.e. p is $\frac{1}{2}$ in the absence of linkage, the efficiency of the statistic in question is 1. It is thus fully efficient for the detection of linkage, and so the use of χ^2 for this purpose is justified. Where the linkage value is small the efficiency of this particular statistic is very low and will lead to the most misleading results. With no recombination, $p = 0$, the efficiency is zero. As an example of the trouble which the use of inefficient statistics leads to, we may consider the estimation of the recombination fraction in some F_2 data showing tight linkage of two factors.

Ex. 9. The data are on the segregation of the two factors, **G**,**g** and **L**,**l** in an F_2 of *Primula sinensis* (De Winton and Haldane, 1936). The factors were in the coupling phase and the following segregation was observed :

TABLE 14

	GL	Gl	gL	gl	Total
Observed .	977	16	19	360	1,372
Expected .	$\frac{n}{4}(2 + P)$	$\frac{n}{4}(1 - P)$	$\frac{n}{4}(1 - P)$	$\frac{n}{4}P$	

where $P = (1 - p)^2$.

Estimation by the method of maximum likelihood as in the previous example leads to the results

$P = 0{\cdot}9507 \quad p = 2{\cdot}50$ per cent assuming $p_1 = p_2$

$V_P = 0{\cdot}3294 \times 10^{-4}$

Estimation by the use of the linear function related to χ^2 leads to the values

$$P = 0{\cdot}9993 \qquad p = 0{\cdot}04 \text{ per cent}$$
$$V_P = 0{\cdot}5469 \times 10^{-3}$$

The efficiency of this last estimate is thus $\dfrac{0{\cdot}00003294}{0{\cdot}0005469}$ or 6·02 per cent.

To express this in another way, an equally precise estimate could have been obtained from eighty-three plants if the method of maximum likelihood had been employed. Nearly 94 per cent of the information in the data has been wasted by inefficient estimation. Furthermore, it will be seen that the estimate obtained is very different from that yielded by the method of maximum likelihood. The difference between the two estimates is significant. Thus the second statistic is not only wasteful of the data but is also clearly wrong.

These deficiencies of the inefficient statistic also result in another serious difficulty. Having obtained an estimate of P we can calculate the expectations of the various classes and apply a χ^2 test to determine whether the whole of the discrepancy originally detected in the data is accounted for by the presence of linkage between the two factors. In the present example the substitution of the value 0·9507, the maximum likelihood estimate, for P leads us to expect the four classes in the frequencies shown in Table 15, middle row.

TABLE 15

		GL	Gl	gL	gl	Total
Observed .		977	16	19	360	1,372
Expected	$P = 0{\cdot}9507$	1012·09	16·91	16·91	326·09	1,372
	$P = 0{\cdot}9993$	1028·76	0·24	0·24	342·76	1,372

Calculating χ^2 for the agreement of these expected values and the observed gives the value $\chi^2 = 5{\cdot}050$. The number of degrees of freedom is reduced from three, the number when testing goodness of fit on the assumption of no linkage, to two because a statistic has been calculated and so the number of classes which can arbitrarily be filled is one less than before. This value of χ^2 for two degrees of freedom has a probability of between 0·10 and 0·05, and so does not indicate any serious deviation from expectation. Nearly the whole of the original discrepancy, from the hypothesis of good single factor ratios and no linkage, is accounted for by the assumption of linkage, as an analysis of χ^2 would have suggested.

If we take the second estimate of P, 0·9993, and calculate the expectation from it the frequencies are very different. In this case it is not possible to calculate a χ^2 for the goodness of fit as the expectations are very much less than 5. It will be remembered that χ^2 cannot be used in such cases, as it fails to follow the tabulated distribution at all closely, when this minimum expectation is not reached. It is, however, quite clear that in the present example the fit of the observed values to the second set of expectations, those derived by inefficient estimation, is very poor. It can be stated as a general rule that the χ^2 test of goodness of fit cannot be used when some statistic, necessary for the calculation of the expectations, has been arrived at by inefficient estimation (Fisher, 1928).

So far the only efficient method of estimation considered has been that of maximum likelihood. For any given problem of estimation other efficient methods may, and often do, exist. Linkage values are estimated efficiently by the use of the product formula (Fisher and Balmakund, 1928; Immer, 1930). This method equates the fraction $\frac{a_1a_4}{a_2a_3}$ to its

expected value $\frac{m_1 m_4}{m_2 m_3}$ to give the equations of estimation. In the case of the F_2 the equation is

$$\frac{a_1 a_4}{a_2 a_3} = \frac{2P + P^2}{1 - 2P + P^2}$$

Any value of the left side of the equation corresponds to but one value of P, and consequently tables for the solution of these equations are easily made. Such tables have been prepared by Immer (1930) and it is only necessary to find the value of the fraction $\frac{a_1 a_4}{a_2 a_3}$ and then look up the corresponding value of P or p in the table. This method has a number of advantages in the estimation of recombination fractions, particularly that by its use certain difficulties encountered in handling data showing poor viability of certain genotypes are minimized. It will be considered in more detail in this special connexion in a later chapter.

The method of maximum likelihood is, however, the only method which leads to efficient estimates for all types of problems of estimation. All other methods need testing against this method before it is decided that they are efficient and consequently to be used.

CHAPTER VI

INFORMATION AND THE PLANNING OF EXPERIMENTS (II)

16. THE AMOUNT OF INFORMATION AND ITS USES

THE previous chapter has introduced the concept of a finite amount of information concerning a linkage value in a given body of data. Any given segregation will allow of the calculation of a linkage value whose maximum precision is determined by the expected values of the classes and the total number of individuals in the family. The whole of this information relevant to the recombination fraction p is extracted by the use of the method of maximum likelihood, but certain other statistics utilize only part of it, and consequently are less efficient estimates of the parameter in question. This concept of the amount of information present in any body of data is of great value in the planning of linkage experiments and a method has been developed to allow of its exact treatment.

The greater the amount of information concerning the recombination fraction, the greater the precision, or the less the variance, of thc estimate, and so it is convenient to define the total amount of information in the data as the inverse of the variance of that statistic obtained by the use of the method of maximum likelihood.

$$I_p = \frac{1}{V_p}$$

A further convenient distinction may be drawn

between the amount of information yielded by a whole family of n individuals and the average amount yielded by a single individual of the family. Then

$$I_p = ni_p.$$

The variance of the estimate of p obtained by the method of maximum likelihood is always a minimum and is calculated from the formula

$$-\frac{1}{V_p} = nS\left(m\frac{d^2 \log m}{dp^2}\right)$$

This gives an easy method of calculating the value of i_p

$$i_p = -S\left(m\frac{d^2 \log m}{dp^2}\right)$$

which is at a maximum for the body of data in question. An alternative formula that is sometimes easier and more convenient to use is

$$i_p = S\left[\frac{1}{m}\left(\frac{dm}{dp}\right)^2\right]$$

(*N.B.*—This is identical with the previous formula.)

The calculation of i_p allows of the comparison of the precision of estimates of a parameter from two entirely different types of data. Note that for this purpose we use i_p rather than I_p since the former is independent of the number of individuals in the family. It is a measure of the value of single individuals in segregations of the types under consideration.

Ex. 10. As an example of the calculation and use of quantities of information let us consider the precision of the linkage values obtained from backcross (**AaBb** × **aabb**) and F_2 (**AaBb** × **AaBb**) data (Mather 1936*a*). The segregations expected in terms of p in families of these types have been worked out in the previous chapter. For this purpose we assume, in the case of the F_2, that $p_1 = p_2$ and so obtain an estimate of p from the value of P which is obtainable

from the data. The method of calculating the amount of information per plant from the formulae

$$i_p = S\left[\frac{1}{m}\left(\frac{dm}{dp}\right)^2\right]$$

in the backcross is shown in Table 16.

TABLE 16 (Mather 1936*a*)

Class	Coupling			Repulsion		
	m	$\frac{dm}{dp}$	i	m	$\frac{dm}{dp}$	i_p
AB/ab	$\frac{1}{2}(1-p)$	$-\frac{1}{2}$	$\frac{1}{2(1-p)}$	$\frac{1}{2}p$	$\frac{1}{2}$	$\frac{1}{2p}$
Ab/ab	$\frac{1}{2}p$	$\frac{1}{2}$	$\frac{1}{2p}$	$\frac{1}{2}(1-p)$	$-\frac{1}{2}$	$\frac{1}{2(1-p)}$
aB/ab	$\frac{1}{2}p$	$\frac{1}{2}$	$\frac{1}{2p}$	$\frac{1}{2}(1-p)$	$-\frac{1}{2}$	$\frac{1}{2(1-p)}$
ab/ab	$\frac{1}{2}(1-p)$	$-\frac{1}{2}$	$\frac{1}{2(1-p)}$	$\frac{1}{2}p$	$\frac{1}{2}$	$\frac{1}{2p}$
Total .	1	0	$\frac{1}{p(1-p)}$	1	0	$\frac{1}{p(1-p)}$

The first column gives the class genotype and it should be noted that each genotype is, in the usual case, distinguishable phenotypically. The second column gives the expectation of each class in terms of p and the third gives the first differential of the expectation with respect to p. From the values in the second and third columns it is easy to calculate the amount of information as shown in the fourth column. These class amounts are summed and give the value of i_p, the average amount of information per individual in a backcross family. The coupling and repulsion phases are considered separately, but it will be seen that, as might be expected, they eventually give the same value for i_p.

The value of i_p from F_2 data depends on how completely the family is classified. We may conveniently recognize three types of classification. First of all, classification may be complete. This will, even under the most favourable circumstances, involve the separation of the doubly heterozygous class (**AaBb**) into those in the coupling and repulsion phases by the use of progeny tests. If this is done the classes are ten in number, with the expectations and amounts of information concerning p as shown in Table 17. This table is laid out precisely as the previous one. Note that again coupling and repulsion yield the same values for i_p. It will be seen that a completely classified F_2 gives twice as much information about p as does a backcross, which is not surprising when it is remembered that an F_2 gives information about recombination in both male and female gametogenesis.

TABLE 17 (Mather 1936*a*)

Class	Coupling			Repulsion		
	m	$\frac{dm}{dp}$	i_p	m	$\frac{dm}{dp}$	i_p
AB/AB	$\frac{1}{4}(1-p)^2$	$-\frac{1}{2}(1-p)$	1	$\frac{1}{4}p^2$	$\frac{1}{2}p$	1
AB/Ab	$\frac{1}{2}p(1-p)$	$\frac{1}{2}(1-2p)$	$\frac{\frac{1}{2}(1-2p)^2}{p(1-p)}$	$\frac{1}{2}p(1-p)$	$\frac{1}{2}(1-2p)$	$\frac{\frac{1}{2}(1-2p)^2}{p(1-p)}$
AB/aB	$\frac{1}{2}p(1-p)$	$\frac{1}{2}(1-2p)$	$\frac{\frac{1}{2}(1-2p)^2}{p(1-p)}$	$\frac{1}{2}p(1-p)$	$\frac{1}{2}(1-2p)$	$\frac{\frac{1}{2}(1-2p)^2}{p(1-p)}$
AB/ab	$\frac{1}{2}(1-p)^2$	$-(1-p)$	2	$\frac{1}{2}p^2$	p	2
Ab/aB	$\frac{1}{2}p^2$	p	2	$\frac{1}{2}(1-p)^2$	$-(1-p)$	2
Ab/Ab	$\frac{1}{4}p^2$	$\frac{1}{2}p$	1	$\frac{1}{4}(1-p)^2$	$-\frac{1}{2}(1-p)$	1
Ab/ab	$\frac{1}{2}p(1-p)$	$\frac{1}{2}(1-2p)$	$\frac{\frac{1}{2}(1-2p)^2}{p(1-p)}$	$\frac{1}{2}p(1-p)$	$\frac{1}{2}(1-2p)$	$\frac{\frac{1}{2}(1-2p)^2}{p(1-p)}$
aB/aB	$\frac{1}{4}p^2$	$\frac{1}{2}p$	1	$\frac{1}{4}(1-p)^2$	$-\frac{1}{2}(1-p)$	1
aB/ab	$\frac{1}{2}p(1-p)$	$\frac{1}{2}(1-2p)$	$\frac{\frac{1}{2}(1-2p)^2}{p(1-p)}$	$\frac{1}{2}p(1-p)$	$\frac{1}{2}(1-2p)$	$\frac{\frac{1}{2}(1-2p)^2}{p(1-p)}$
ab/ab	$\frac{1}{4}(1-p)^2$	$-\frac{1}{2}(1-p)$	1	$\frac{1}{4}p^2$	$\frac{1}{2}p$	1
	1	0	$\frac{2}{p(1-p)}$	1	0	$\frac{2}{p(1-p)}$

The second type of classification that it is convenient to consider is that which arises when factors **A,a** and **B,b** both show incomplete dominance, i.e. when the three possible combinations of the two allelomorphs at one locus are recognizable phenotypically. It is assumed that classification for one factor is independent of the allelomorphs present at the other locus, i.e. that the two factors are independent in expression. Such classification will give nine phenotypic classes of which one, **AaBb**, will comprise two distinct genotypes, the coupling and repulsion heterozygotes. This class will be composite and have the expectation $\frac{1}{2}(1 - 2p + 2p^2)$ which contributes $\frac{2(1-2p)^2}{1-2p+2p^2}$ to the value of i_p. This contribution replaces those of 2 and 2 which are made by the separated coupling and repulsion double heterozygotes. Hence the value of i_p obtained from an F_2 classified in this manner is :

$$\frac{2}{p(1-p)} - 4 + \frac{2(1-2p)^2}{1-2p+2p^2}$$

or

$$\frac{2(1-3p+3p^2)}{p(1-p)(1-2p+2p^2)}$$

Finally we may consider the commonest case of all, that of complete dominance of **A** over its allelomorph **a** and of **B** over **b**. There are then four phenotypic classes in the F_2 of which three contain more than one genotypic class. The values of the contributions of these classes to i_p are worked out in Table 18.

It will be seen that coupling and repulsion F_2 do not yield equal amounts of information concerning p when classification is of this type.

We can now compare the efficiencies of the backcross and F_2 of varying degrees of classification, for the calculation of linkage values. For this purpose it is best to take the amount of information given by

TABLE 18 (Mather 1936*a*)

Class	Coupling			Repulsion		
	m	$\frac{dm}{dp}$	i_p	m	$\frac{dm}{dp}$	i_p
AB	$\frac{1}{4}(3-2p+p^2)$	$\frac{-(1-p)}{2}$	$\frac{(1-p)^2}{3-2p+p^2}$	$\frac{1}{4}(2+p^2)$	$\frac{1}{2}p$	$\frac{p^2}{2+p^2}$
Ab } **aB** }	$\frac{1}{2}(2p-p^2)$	$1-p$	$\frac{2(1-p)^2}{2p-p^2}$	$\frac{1}{2}(1-p^2)$	$-p$	$\frac{2p^2}{1-p^2}$
ab	$\frac{1}{4}(1-2p+p^2)$	$\frac{-(1-p)}{2}$	1	$\frac{1}{4}p^2$	$\frac{1}{2}p$	1
	1	0	$\frac{2(3-4p+2p^2)}{p(2-p)(3-2p+p^2)}$	1	0	$\frac{2(1+2p^2)}{(2+p^2)(1-p^2)}$

the backcross as standard because in this way infinite values of i_p are avoided. With this standard the relative values of the different types of data are as shown in Table 19.

TABLE 19 (Mather 1936*a*)

Backcross		1
F_2 completely classified . .		2
F_2 incomplete dominance . .		$\frac{2(1-3p+3p^2)}{1-2p+2p^2}$
F_2 complete dominance	Coupling	$\frac{2(1-p)(3-4p+2p^2)}{(2-p)(3-2p+p^2)}$
	Repulsion	$\frac{2p(1+2p^2)}{(2+p^2)(1+p)}$

These relative values are dependent on p itself. Consequently a clear idea of the meaning of these values will be obtained by plotting the value against p in the form of a graph. For this purpose repulsion is considered to be an extension of coupling. Thus $p = 0{\cdot}3$ in repulsion may be plotted as $p = 0{\cdot}7$ in coupling. Fig. 1 is then obtained.

This Fig. 1 is instructive in a number of ways. In the first place it is easy to see that in the case of

two incompletely dominant factors the F_2 is as good as the backcross for the detection of linkage, i.e. the detection of deviation from $p = 0{\cdot}5$. For the measurement of linkage values, particularly when the recombination value is small in either phase, this F_2 is better than the backcross in that it gives more information about p and so a more precise

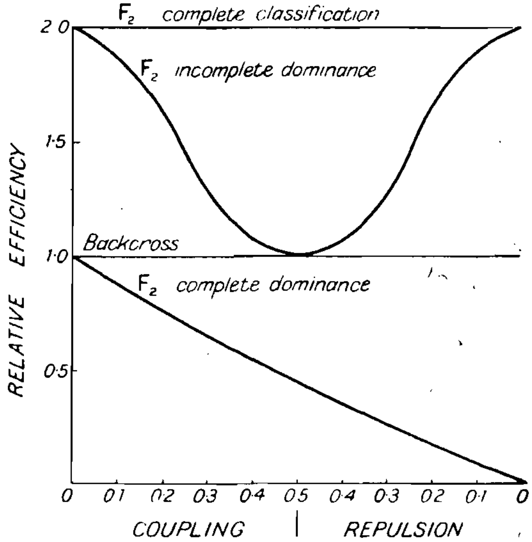

FIG. 1 (after Mather 1936*a*)

estimate of the recombination fraction. The case of factors having one allelomorph completely dominant over the other is, however, very different. In close coupling the F_2 is almost as good as the backcross, but in close repulsion this is far from the case. The backcross is then vastly better for the estimation of p. Where there is no linkage, a case important in that this is the hypothesis tested in order to detect linkage, the F_2 has 4/9 of the value of the backcross.

It is thus not so good for the detection of loose linkages. In general an F_2 of this the most usual type is much less efficient than the backcross except for the single case of close coupling. So it is not to be recommended, when the alternatives are of equal practical ease.

In this way, given the behaviour of the two single factors, it is possible to decide on (*a*) the best type of family for the detection of linkage and (*b*) the best type of family for its estimation. Practical considerations also enter into the question, e.g. the ease of backcrossing as compared with inbreeding is an important consideration in plant genetics. These considerations may to some extent set off the statistical advantages of a given type of data, but the experimenter, knowing his crop, will be able to form a fairly accurate estimate of the relative importance of the various considerations and will be able to reach a confident conclusion as to the best method of tackling the particular problem at hand.

Fig. 1 also illustrates another very important point, that of the loss of information resulting from incomplete classification. Any F_2 contains twice as much information about the recombination fraction as a backcross, but the limitations of classification result in a certain loss, which, in the case of completely dominant genes in close repulsion, may amount to an extremely large proportion of the whole. It is very clear that data should be as completely classified as is immediately possible in order to reduce this loss to a minimum. Where the further classification involves progeny tests, as it would in F_2 families, the number of plants and the labour may or may not be such as to render the extra classification unprofitable as compared with growing further families of a similar kind. The policy to be adopted with respect to growing F_3s to test the F_2 individual genotypes or growing further F_2 families is capable of exact treatment by the calculation of quantities

of information. This has been done in some detail by Immer (1934) and Mather (1936*a*) and need not be treated here. It is sufficient to say that with a repulsion F_2 it is profitable to test the genotypes of the singly dominant classes (**Ab** and **aB**) when p is less than 0·08, and to test the genotypes of the doubly dominant class (**AB**) when p is less than 0·22. Thus the use of this method allows of specification of the classes which can be tested with profit. It allows of very precise planning of such experiments.

Ex. 11. The previous example was a consideration of a relatively familiar problem, but one of the great advantages of the method of approach developed above is that it helps to clarify policy when unusual genetical situations are encountered. Reference to various papers will usually provide all the information and experience necessary for reaching a decision in connexion with the more ordinary genetical situations, but this is not true of some other less usual circumstances. The experimenter is then forced to deal with the situation unaided by previous experience.

As an example let us consider the case of linkage between one gene and another which is a member of a pair of complementary factors (Hutchinson 1929). If these genes are respectively **A,a** and **B,b**, then the expression of the **B,b** difference is dependent on the presence of one or other of the two allelomorphs of a third factor, **C,c**. More precisely the genotypes **Bcc**, **bbC** and **bbcc** are phenotypically alike. In order to measure the linkage between **A,a** and **B,b** it is, of course, necessary to raise the double heterozygote **AaBb**, It will usually be the case that the individual is also heterozygous for **C,c** as it is not easy to tell which of the twc complementary factors is involved in the linkage. When the heterozygote is of this type, **AaBbCc**, several possible crosses are open to the experimenter. He may cross to a stock recessive for **a** and for one of the complementary

factors while being homozygous dominant for the other complementary.

There will be two such crosses possible, **AaBbCc** × **aabbCC**, and **AaBbCc** × **aaBBcc.** The former will yield extremely good data from which the linkage can be estimated. The latter will show no immediate segregation of **B** and **b** and so will be useless or nearly so, from the point of view of linkage estimation. If it is not known which of the two complementaries is linked to **A,a** the probability of any cross of this type being that which will give useful linkage data is $\frac{1}{2}$.

Another line of policy is to cross the triple heterozygote to a triple recessive, **AaBbCc** × **aabbcc.** The classification for **B** and **b** will be incomplete owing to the segregation of **C** and **c**. The classification for **A** and **a** is, however, complete.

A third possibility is that of selfing or inbreeding the triple heterozygotes. Classification for **A** and **a** will be incomplete, but there will be less disturbance of the **B,b** classification as only $\frac{1}{4}$ of the progeny will have **cc**, which renders the **B,b** difference undetectable.

There are other possibilities, of course, but these seem to be the most likely to arise in practice. The problem then resolves itself into one of choosing which of three types of cross to use when each type has its peculiar disadvantages. Assume that p is the same in male and female gametogenesis and then it can be shown that the expectations of the four scorable phenotypic classes relevant to the linkage in the three types of cross are:

LIBRARY
TNAU, Coimbatore - 3
000051063

b(C) **A(c)**	**aB(C)**	**ab(C)** + **a(c)**
$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$\frac{1}{4}p$	$\frac{1}{4}(1+p)$	$\frac{1}{4}(2-p)$
$-3P)$	$\frac{1}{16}(3-3P)$	$\frac{1}{16}(1+3P)$

The expected segregations given above are those for the coupling phase. Similar expressions for linkage in the repulsion phase may be formulated by substituting $1 - p$ for p in these expressions.

Having the expected frequencies for each class in the various types of family we are now in a position to work out the values of i_p in terms of p. These are obtained from the formula $i_p = S\left[\frac{1}{m}\left(\frac{dm}{dp}\right)^2\right]$ as in the previous example. We then find the following values :

$$(1)\begin{cases}\textbf{AaBbCc} \times \textbf{aabbCC} \; i_p = \dfrac{1}{p(1-p)} \\ \textbf{AaBbCc} \times \textbf{aaBBcc} \; i_p = 0\end{cases}$$

(2) **AaBbCc** × **aabbcc**

$$i_p = \frac{1 + 2p - 2p^2}{2p(1-p)(1+p)(2-p)}$$

(3) **AaBbCc** × **AaBbCc**

$$i_p = \frac{3P(5 + 2P - 4P^2)}{(2+P)(1-P)(2-P)(1+3P)}$$

where $P = (1 - p)^2$.

Taking the first type of family, the backcross giving completely classifiable segregation of the linked factors, as standard the following relative values of the families are obtained.

$$(1)\begin{cases}1\\0\end{cases}$$

$$(2) \quad \frac{1 + 2p - 2p^2}{2(1+p)(2-p)}$$

$$(3) \quad \left(\frac{3P(5 + 2P - 4P^2)}{(2+P)(2-P)(1+3P)}\right)\left(\frac{1-p}{2-p}\right)$$

The values of the different amounts of information may be plotted against the value of p, taking repulsion as an extension of coupling, as previously.

Fig. 2 is then obtained. This figure supplies all the information necessary for our purpose. In the first place it is seen that the first type of family is very much more informative than any others. The relative values of the third and fourth types of family, the complete backcross and the F_2, change over the range $p = 0{\cdot}0$ in coupling to $p = 0{\cdot}0$ in repulsion. Sometimes the former is better than the latter, and

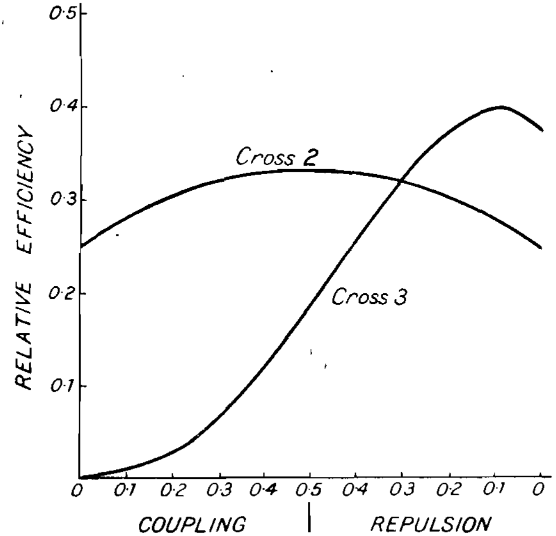

FIG. 2

sometimes the reverse is the case. The important point is, however, that the cross of type 1, **AaBbCc** × **aabbCC**, is always at least 2·5 times as informative as the better of the other types and has usually an even greater value than this. Since it is usually impossible to draw a distinction between crosses of types 1 and 2, owing to the lack of knowledge as to which complementary factor is linked with **Aa**, we

must discount the value of cross 1 by half its value. It is clear, however, that, no matter what linkage value is concerned or whether the problem is one of detection rather than estimation, the most profitable policy is to grow equal numbers of progenies from crosses of types 1 and 2, **AaBbCc** × **aabbCC** and **AaBbCc** × **aaBBcc.** One cross will give no immediate information about linkage, but since the other supplies more than twice the information of the better alternative, the total information obtained by this procedure will still be greater than that which can be obtained in any other way. This result is certainly not one that could have been easily foreseen. It is in many ways novel to adopt a policy that involves the deliberate wastage of half the individuals produced. In general such a policy would probably not be the best, but in this particular case the limitations imposed on classification by the factor interaction are such that one type of cross, itself distinguishable from a useless alternative, has such a preponderant value that discarding half the progenies is justified. The value of planning the linkage experiments is here demonstrated in a most striking way.

CHAPTER VII

COMBINED ESTIMATION AND TESTING HETEROGENEITY

17. COMBINED ESTIMATION

ONLY simple problems of estimation have been considered up to the present. These have consisted of estimation from single families of given types, in which connexion it has been shown that the method of maximum likelihood has a number of advantages. This method of estimation is also of value in the solution of two somewhat different problems, viz. those of arriving at the best estimate of a parameter when data of several different kinds are available, and of testing the homogeneity of such aggregates of data (Mather 1935).

Let us first consider the question of combined estimation. This is well illustrated by the estimation of the simplex index of separation in autotetraploid segregations. In organisms of this type there is no simple expectation for single factor segregations and, in order to describe the segregation of a factor, it is necessary to calculate the value of the parameter named above (*see* Mather, 1936*b*). The expected gametic segregation of a simplex autotetraploid, i.e. one whose constitution is **Aaaa**, is **A**(**AA** or **Aa**) $\frac{1}{8}(4 - \alpha)$: **aa** $\frac{1}{8}(4 + \alpha)$ where α is the simplex index of separation. On selfing, one expects a segregation of

$$\mathbf{A}\ \tfrac{1}{64}(48 - 8\alpha - \alpha^2) : \mathbf{a}\ \tfrac{1}{64}(16 + 8\alpha + \alpha^2)$$

Ex. 12. Sansome (quoted by Mather 1936*b*) has

observed the following segregations for the factor **R,r** in simplex autotetraploids of the tomato.

TABLE 20

	R	r	Total
(**Aaaa** × **aaaa**) Backcross	48	65	113
(**Aaaa** × **Aaaa**) F_2 .	605	221	826

The former is clearly the gametic segregation of the simplex individual and the latter is the segregation to be obtained from selfing the same plant. Hence the expectations are, as given above.

TABLE 21

	R	r	Total
Backcross .	$\frac{1}{8}(4 - \alpha)$	$\frac{1}{8}(4 + \alpha)$	1
F_2 . . .	$\frac{1}{64}(48 - 8\alpha - \alpha^2)$	$\frac{1}{64}(16 + 8\alpha + \alpha^2)$	1

What is the best estimate of α that can be obtained from these data ?

The likelihoods of obtaining such segregations separately are, following the argument given in Chapter V,

Backcross $C_1[\frac{1}{8}(4 - \alpha)]^{48}[\frac{1}{8}(4 + \alpha)]^{65}$

F_2 . . $C_2[\frac{1}{64}(48-8\alpha-\alpha^2)]^{605}[\frac{1}{64}(16+8\alpha+\alpha^2)]^{221}$

The likelihood of obtaining these two segregations jointly is the product of the two individual likelihoods. Then the logarithm of the joint likelihood will be given by the sum of the individual logarithm likelihood expressions, i.e. will be :

$$L = 48 \log (4 - \alpha) + 65 \log (4 + \alpha) + 605 \log (48 - 8\alpha - \alpha^2) + 221 \log (16 + 8\alpha + \alpha^2)$$

The maximum likelihood estimate of α will be obtained by maximizing this summed logarithm likelihood expression with respect to α. Differentiating and equating to zero we obtain :

$$\frac{dL}{d\alpha} = -\frac{48}{4 - \alpha} + \frac{65}{4 + \alpha} - \frac{605(8 + 2\alpha)}{48 - 8\alpha - \alpha^2} + \frac{221(8 + 2\alpha)}{16 + 8\alpha + \alpha^2} = 0$$

The solution of this equation is not difficult to arrive at by algebraic methods, but it will serve as an example of the alternative process of solution by arithmetic trial and interpolation. This latter method is of great value where there exists no easy algebraic method of approach.

As a first approximation to the solution of the equation put $\alpha = 0{\cdot}20$. The value of the left side of the equation may then be calculated as shown in Table 22. It will be seen that this value is negative and so we have chosen too high a value for α. Then repeat the calculation using $\alpha = 0{\cdot}10$. This value is clearly too low for α. We then make a linear interpolation between these values of α. Thus

$$\alpha = 0{\cdot}10 + \frac{6{\cdot}2227}{6{\cdot}2227 + 1{\cdot}5377} = 0{\cdot}10 + 0{\cdot}080 = 0{\cdot}180.$$

Trial of this value shows it to be a trifle too small, a small positive value being found for the expression. On trying 0·181 we obtain a negative value for the left side of the equation and so can again interpolate between 0·18 and 0·181. This gives us 0·1803 as a third approximation to the true value of α. The value arrived at by algebraic solution of the equations agrees with this value to four decimal

TABLE 22

α	0·20	0·10	0·18	0·181
$-\dfrac{48}{4-\alpha}$	− 12·6316	− 12·3077	− 12·5654	− 12·5687
$\dfrac{65}{4+\alpha}$	15·4762	15·8537	15·5502	15·5465
$-\dfrac{605(8+2\alpha)}{48-8\alpha-\alpha^2}$	− *109·6204*	− *105·1282*	− *108·7054*	− *108·7509*
$\dfrac{221(8+2\alpha)}{16+8\alpha+\alpha^2}$	105·2381	107·8049	105·7416	105·7163
Total . . .	− 1·5377	6·2227	0·0210	− 0·0568

places. It may be noted that although the first interpolation was somewhat inaccurate owing to the wrong assumption of a straight line relation between the values of α and the expression, the second interpolation was more accurate. In general, the closer the limits between which interpolation is made, the more accurate the result. Further interpolation could be tried to obtain the value of α even more accurately; but not more than four decimal places are warranted by the data in this case and so further calculation would be wasted.

The method of arithmetic approximation has another great advantage in that it automatically leads to an estimate of the variance of α. It will be remembered that I_α, the inverse of V_α, can be obtained from the second differential of the logarithm likelihood expression with respect to α. In other words, I_α is *the rate of change on α of the maximum likelihood* expression, which is itself the first differential with respect to α. Now when

$$\alpha = 0{\cdot}180 \quad \frac{dL}{d\alpha} = 0{\cdot}0210$$

and when $\alpha = 0{\cdot}181 \quad \frac{dL}{d\alpha} = -\,0{\cdot}0568$

Hence a change of 0·001 in α results in a change of 0·0778 in the value of $\frac{dL}{d\alpha}$. Hence the rate of change of $\frac{dL}{d\alpha}$ on α in this region is $\frac{0{\cdot}0778}{0{\cdot}001}$ or, in other words, $I_\alpha = 77{\cdot}8$.

If we calculated I_α from the difference of the maximum likelihood values when $\alpha = 0{\cdot}10$ and $0{\cdot}20$, the result would be somewhat smaller than that obtained above because the rate of change decreases as the approximation to the value of α becomes coarser. Consequently, the closest approximations to the true value should be used in applying this empirical method of obtaining quantities of information.

Having obtained $I_\alpha = 77{\cdot}8$ we find that $V_\alpha = 0{\cdot}01284$ and that $s_\alpha = \sqrt{V_\alpha} = 0{\cdot}113$.

It will be noted that the value of I_α obtained in this way is not precisely the same as that obtained from the use of the formulae

$$I_\alpha = nS\left[\frac{1}{m}\left(\frac{d^2m}{d\alpha}\right)^2\right] = -\ nS\left(m\frac{d^2 \log m}{d\alpha^2}\right)$$

The value obtained arithmetically is the actual amount of information present in this particular body of data. The value yielded by the formulae quoted is the *mean* amount of information to be expected from a large number of families of this kind and size. In the present case the mean value of I_α is 78·1 or slightly more than the value obtained arithmetically. Either value may be used for the purpose of estimating the variance of the parameter as they never differ by much; but the expected mean amount of information should always be employed in planning experiments as shown in the last chapter.

18. TESTING HETEROGENEITY

The other type of problem to which the method of maximum likelihood is adapted is that of the detection of heterogeneity between different bodies of data concerning the same parameter. Its use in this respect may also be illustrated by example (Mather 1935).

Ex. 13. Bateson (1909) records segregations for the two genes purple-red flower colour and long-round pollen in the Sweet Pea. A family showed the following segregation :

Purple Long	Purple Round	Red Long	Red Round
296	19	27	85

indicating linkage in the coupling phase. An F_3 also gave evidence of linkage in coupling in the segregation

Purple Long	Purple Round	Red Long	Red Round
583	26	24	170

Do these two sets of data agree in showing the same recombination fraction for the two genes ?

This question may be approached in several ways. First, we may calculate the value of $P(= (1 - p)^2)$ and its variance for each set of data separately and test the significance of the difference of the linkage values by using the formula $\frac{d}{s} = \frac{P_1 - P_2}{\sqrt{V_{P_1} + V_{P_2}}}$. The numerator of this fraction is the difference between the two values of P and the denominator is the estimated standard error of this difference. Using the methods of Chapter V we find

$$d = P_1 - P_2 \qquad = 0{\cdot}088636$$

$$s = \sqrt{V_{P_1} + V_{P_2}} = 0{\cdot}033786$$

and the difference divided by its standard error is 2·6235.

A table of probabilities of normal deviates (Table 1) shows this value to have a probability of just less than 1 per cent. This value also corresponds to a χ^2 of 6·8828, which is obtained by squaring the value 2·6235. This method of analysis suffers from a serious disadvantage. The two variances are calculated on the bases of the two separate estimates of P. We desire to test the hypothesis that these data agree in showing one value of P, so the variances should have been calculated on the basis of the best combined estimate of P. Hence the values of the variances reached above are not correct. In fact, one is too large and the other too small. The two discrepancies, though of opposite sign, are not of necessity equal in magnitude and need not balance.

Thus we are led to approach the problem as one of combined estimation. Assuming homogeneity, we can add the two sets of data together and estimate P from the totals, obtaining 0·843047. This value may then be used in the formulation of expected segregation for the two families. From the relations of

these expectations to the observed segregations we may calculate χ^2 for the F_2 and F_3 families separately as is done in Table 23. In this table the two central or singly dominant classes of the family are added together for purposes of estimation as they have the same expectation in terms of P. Any difference between them will be dependent on the single factor ratios alone, and so holds no interest for us in the discussion of this present problem.

TABLE 23 (Mather 1935)

Observed (a)	Expected (mn) [P = 0·843047]	$\frac{dmn}{dP}$	Discrepancy in Likelihood Equation $\left[\frac{a}{m}\frac{dm}{dP}\right]$	Amount of Information $n\left[\frac{1}{m}\left(\frac{dm}{dP}\right)^2\right]$	χ^2 $\frac{(a-nm)^2}{nm}$
F_2 296	303·495	106·75	104·114	37·55	0·1851
46	33·510	− 213·50	− 293·077	1360·26	4·6553
85	89·995	106·75	100·825	126·62	0·2772
Total 427	427·000	0·00	− 88·138	1524·43	
F_3 583	570·742	200·75	205·062	70·61	0·2633
50	63·016	− 401·50	− 318·570	2558·12	2·6885
170	169·242	200·75	201·649	238·12	0·0034
Total 803	803·000	0·00	+ 88·141	2866·85	8·0728

The contributions to χ^2 of the two sets of families are 5·1176 and 2·9552, giving a total of 8·0728. This will have three degrees of freedom because there were two degrees of freedom in each family (three classes), but one of the total of four has been lost in estimating the linkage value. Consequently we are led to the conclusion that the deviation of the two families from their expectation is just significant, the probability of χ^2 being just less than 5 per cent.

The test may be made more sensitive, however, since at present it concerns not only the agreement of the two families with respect to the linkage value,

but also two superfluous degrees of freedom concerning single factor ratios in which we are not particularly interested. These superfluous portions of χ^2 may be removed as follows. The two distinct values of P yielded by the families separately have already been found. Expectations based on these may be formulated and agreement between them and the corresponding observations may be tested by calculating χ^2 values. We find these two χ^2 values to be 0·0228 and 0·2397 (*see* Table 24). If these values are subtracted from the χ^2 calculated on the basis of the joint estimate of P, the remainder is concerned solely with difference between the families and has one degree of freedom. Since the single factor ratios of both genes in both families are good, the discrepancy, if any, between the families must be due to discrepancies in the values of P shown by the two segregations. The remainder of χ^2 obtained in this way is 7·8101 for one degree of freedom and is highly significant. The two sets of data do not agree in the recombination fractions that they show.

TABLE 24

	Observed (a)	Expected (nm)	Deviation ($a - nm$)	χ^2 $\left[\frac{(a-nm)^2}{nm}\right]$
F_2	296	297·354	− 1·354	0·0062
	46	45·792	+ 0·208	0·0010
	85	83·854	+ 1·146	0·0156
Total	427	427·000	0·000	0·0228
F_3	583	576·987	+ 6·013	0·0627
	50	50·526	− 0·526	0·0055
	170	175·487	− 5·487	0·1717
	803	803·000	0·000	0·2399

The χ^2 obtained in this way is much more significant than that obtained by the first method discussed.

This difference well exemplifies the advantages of using an exact method of approach. Differences are likely to be overlooked if an insensitive test is applied.

The exact method of calculating the heterogeneity χ^2s is very long but may be shortened considerably by using the formula

$$\chi^2 = Q^2 + \frac{1}{I_P}\left[\frac{a_1}{2+P} - \frac{a_2+a_3}{1-P} + \frac{a_4}{P}\right]^2$$

where P is the joint estimate of the function of the recombination fraction. Q^2 is that portion of χ^2 which is accounted for by the deviations of the single factor ratios from their expectations (Fisher, 1949*a*). Consequently the portion $\frac{1}{I_P}\left[\frac{a_1}{2+P} - \frac{a_2+a_3}{1-P} + \frac{a_4}{P}\right]$ is a χ^2 which is dependent on the discrepancy between the calculated joint value of P and the value afforded by the body of data in question. It is obtained by squaring the deviation from zero of the maximum likelihood expression and dividing by the amount of information concerning P in the particular body of data. Thus the χ^2 for heterogeneity of the linkage values is given by the formula

$$\chi^2 = S\left[\frac{1}{I_P}\left\{\frac{a_1}{2+P} - \frac{a_2+a_3}{1-P} + \frac{a_4}{P}\right\}^2\right]$$

which may be written $\chi^2 = S\left(\frac{D^2}{I}\right)$

summation proceeding over all bodies of data.

The calculation of the heterogeneity in the example under consideration is shown in Table 23. The first column shows the observed segregations. In the second column is the value expected. The third column is found from the differential of the expected frequencies. Thus for the first row of the table $m = \frac{n}{4}(2+P)$ and so $\frac{dm}{dP} = \frac{n}{4}$. The fourth column, the discrepancy in the likelihood expression, is found

arithmetically row by row by dividing the product of the values in columns one and three by the value in column two. The fifth column, the contribution to the amount of information, is obtained by squaring the value in column three and dividing by that in column two. The deviations of the two families from the maximum likelihood solution is found by summation in column four. It will be seen that the two deviations are opposite in sign and almost equal, so showing that the combined value of P has been properly estimated. The amount of information from each family is found by similar summation in column five. Then the contribution of each family to χ^2 is found by squaring the family total in column four, the deviation from the maximum likelihood solution, and dividing by the family amount of information from column five. The two contributions obtained in this way are 5·0959 and 2·7099, giving a χ^2 of 7·8058 for one degree of freedom. This is very close to the value obtained by the previous method of calculation. Thus the conclusion that the families do not agree in the values of P that they show may be arrived at conveniently and quickly by this method, which involves only the estimation of the best joint-statistic of P.

Ex. 14. As a more complex example of the use of combined estimation in arriving at joint estimates and in testing heterogeneity we may consider the data of Jenkins (1927) on recombination between the factors **Y**,**y** and **Wx**,**wx** in maize. This author has three types of family which give information about the recombination between these genes, as shown in Table 25.

The expectations for the various classes in each family are shown underneath the observed numbers.

We are concerned to know whether the two single factor ratios are in keeping with Mendelian expectation, what the best estimate of the recombination

TABLE 25

Cross	Y Wx	Ywx	y Wx	y wx	Total
Backcross coupling .	397	297	289	412	1,395
	$\frac{n}{2}(1-p)$	$\frac{n}{2}p$	$\frac{n}{2}p$	$\frac{n}{2}(1-p)$	
Backcross repulsion .	78	136	120	80	414
	$\frac{n}{2}p$	$\frac{n}{2}(1-p)$	$\frac{n}{2}(1-p)$	$\frac{n}{2}p$	
Single backcross repulsion .	461	161	515	130	1,267
	$\frac{n}{4}(1+p)$	$\frac{n}{4}(1-p)$	$\frac{n}{4}(2-p)$	$\frac{n}{4}p$	

fraction is, and whether the data are homogeneous for the recombination fraction.

In each family, segregation for the **Y,y** factor is that of a backcross with expectation of 1 : 1. The analysis of this segregation is then easy and is done by the methods developed in Chapter II. The following analysis of χ^2 is obtained :

	χ^2	D.F.	P.
Deviation . . .	0·0832	1	0·80 – 0·70
Heterogeneity . .	0·8428	2	0·70 – 0·50
Total	0·9260	3	

The case of the factor **Wx,wx** is not, however, so simple. The first two sets of data are both backcrosses for this factor and may be expected to show a 1 : 1 segregation. The third set of data is, on the other hand, an F_2 for this factor and will show a 3 : 1 segregation. How may the joint deviation from expectation be tested ?

A method for doing this test has been developed by Mather (1937). It is based on joint estimation by the method of maximum likelihood.

Let the frequency of gametes carrying the recessive allelomorph, **wx**, be x. Then the gametic output of a heterozygous, **Wx,wx**, plant will be **Wx** $1 - x$: **wx** x. On backcrossing, the observed segre-

gation will be the same as the gametic segregation of the heterozygote. On selfing a heterozygote we shall obtain a segregation of **Wx** $1 - x^2$: **wx** x^2. The observed frequencies are in the backcrosses jointly, **Wx** 884 : **wx** 925 and in the F_2 **Wx** 976 : **wx** 291. The joint logarithm likelihood expression is $L = 884 \log (1 - x) + 925 \log x + 976 \log (1 - x^2) + 291 \log x^2$
and the equation for the joint estimation of x,

$$\frac{dL}{dx} = -\frac{884}{1-x} + \frac{925}{x} - \frac{976 \times 2x}{1-x^2} + \frac{291 \times 2x}{x^2} = 0$$

We are, however, at present solely concerned with the significance of the deviation of the segregations from the simple Mendelian estimation of $x = \frac{1}{2}$. This value for x may be substituted in the maximum likelihood equation and the deviation of the data from the maximum likelihood equation obtained. It is found to be

$$2(925 - 884) - \frac{4}{3}(976 - 3 \times 291)$$

i.e. $-55{\cdot}\dot{3}$.

We next require the amount of information about x yielded jointly by the joint data. The methods of Chapter VI lead us to expect that the backcross will give $\frac{n}{x(1-x)}$ units of information about the value of x and that the F_2 will similarly give $\frac{4n}{1-x^2}$ units of information. Then substituting the expected value of $x = \frac{1}{2}$ the total information becomes $4(884 + 925) + \frac{16}{3}(976 + 291)$ i.e. $139931{\cdot}\dot{3}$. Thus we now have both the deviation of the maximum likelihood expression from the expected answer and the amount of information about the parameter x. We can calculate a χ^2 value from these results by the use of the formula $\chi^2 = \frac{D^2}{I}$ where D is the joint deviation from zero and I the joint information.

This is found to give $\chi^2 = 0{\cdot}2188$ and will clearly

have one degree of freedom, as its value can be made indefinitely small by the adjustment of statistic x. (*N.B.*—This is a deviation, not a heterogeneity, χ^2 as we have not used an estimated value of x.)

The analysis of χ^2 for this factor may now be completed in the usual manner. Each family provides a value of χ^2 calculated by the formulae given in Chapter II. The sum of these three χ^2s is found to be 3·9529, for, of course, three degrees of freedom. From this total the χ^2 for the joint deviation from the Mendelian expectation as calculated above may be deducted to leave a remainder for heterogeneity between the families. The complete analysis is:

	χ^2	D.F.	P.
Deviation . .	0·2188	1	0·70 – 0·50
Heterogeneity . .	3·7341	2	0·20 – 0·10
Total . . .	3·9529	3	

It is clear from the χ^2 analyses for the two single factor ratios that these segregations are quite in keeping with simple expectation and are, also, failing to show any signs of inhomogeneity. So we may proceed to the estimation of the recombination fraction.

Table 25 gives the observed segregations and also the expectations of the various classes in terms of the recombination fraction p. It is then easy to write down the joint logarithm likelihood expression for the three sets of data. In each set of data those classes with the same expectations in terms of p are added together.

$$L = 809 \log (1 - p) + 586 \log p + 256 \log (1 - p) + 158 \log p + 461 \log (1 + p) + 161 \log (1 - p) + 515 \log (2 - p) + 130 \log p$$

The sources of the terms of this expression are obvious. The first two are from the coupling backcross, the second pair from the repulsion backcross and the last four terms from the single backcross.

Maximizing by differentiation with respect to p gives as the equation of estimation:

$$\frac{dL}{dp} = -\frac{809}{1-p} + \frac{586}{p} - \frac{256}{1-p} + \frac{158}{p} + \frac{461}{1+p}$$

$$-\frac{161}{1-p} - \frac{515}{2-p} + \frac{130}{p} = 0$$

This equation is most easily solved by arithmetic approximation, as in Ex. 12. By this means it is found that the value of p is 0·4162 to four decimal places (Table 26).

There is now left the question of heterogeneity of the linkage data. The heterogeneity χ^2 is easily calculated from the figures already obtained in solving the equation of estimation of p. We first of all substitute 0·4162 for p in the equation of estimation. This enables us to find the deviations (D) from zero of the three separate parts of this equation, i.e., those parts coming from the three different sets of data, when the best joint estimate of p is used. We next calculate the amounts of information (I) about p yielded by the three separate sets of data. This is done arithmetically by determining the rate of change of the corresponding portions of the maximum likelihood expression for unit change in p in the neighbourhood of the parameter's best fitting value (cf. Ex. 12). For example, in the coupling backcross $D = 1407·97693 - 1385·78454 = 22·22839$ and

$$I = [(1408·65385 - 1385·27397) - (1405·27578 - 1387·65009)] \times 100 = 5754·19$$

(*see* Table 26). χ^2 may then be calculated for each set of data from the formula $\chi^2 = S\left(\frac{D^2}{I}\right)$. The results are given in Table 27.

TABLE 27

	D	I	χ^2
Backcross coupling .	22·22839	5754·19	0·0859
,, repulsion .	− 58·88116	1662·71	2·0851
Single backcross . .	36·92212	1657·41	0·8225
	0·26935	9074·31	2·9935

TABLE 26

	$p =$	0·40	0·42	0·416	0·417	0·4162
I	$-\frac{809}{1-p}$	− 1348·33333	− 1394·82759	− 1385·27397	− 1387·65009	− 1385·74854
	$\frac{586}{p}$	1465·00000	1395·23810	1408·65385	1405·27578	1407·97693
II	$-\frac{256}{1-p}$	− 426·66667	− 441·37931	− 438·35616	− 439·10806	− 438·50634
	$\frac{158}{p}$	395·00000	376·19048	379·80769	378·89688	379·62518
III	$\frac{461}{1+p}$	329·28571	324·64789	325·56497	325·33522	325·51899
	$-\frac{161}{1-p}$	− 268·33333	− 277·58621	− 275·68493	− 276·15780	− 275·77938
	$-\frac{515}{2-p}$	−321·87500	− 325·94938	− 325·12626	− 325·33165	− 325·16732
	$\frac{130}{p}$	325·00000	309·52381	312·50000	311·75060	312·34983
	Total . .	149·07738	− 34·14221	2·08519	− 6·98912	0·26935

I From Coupling Backcross
II From Repulsion Backcross
III From Single Backcross

The total of the χ^2 values is 2·9935. This will correspond to two degrees of freedom because, of the three provided by the three families, one is used up in calculating the best fitting value of p. The probability of obtaining as large or larger value of χ^2 by chance is 0·30 – 0·20 and so the data may be considered to be homogeneous.

Thus the three families all agree in showing good single factor segregations and a recombination value of 41·62 ± 1·05 per cent.

19. INCOMPLETE MANIFESTATION OF A CHARACTER

The testing of heterogeneity by combined estimation has another important application, viz. to the testing of hypotheses involving segregations for a character which shows incomplete manifestation. In such cases it is usual to grow F_3 progenies to test certain of the F_2 individuals for the presence of the character which may not have been shown phenotypically. The problem is that of incorporating such test progenies in the analysis. The solution of this problem has been described by Smith (1937).

Ex. 15. The actual example used by Smith is that of ' fired ' in certain wheat varieties. This character is apparently controlled by three unlinked complementary genes, i.e. **ABC** plants will be fired, but **ABc**, **AbC**, **aBC**, **Abc**, **aBc**, **abC**, and **abc** plants will be normal. Then in the F_2 raised from triple heterozygotes, the segregation expected is one of 27 fired to 37 normal. Actually it is found that fired plants may sometimes look normal, and so a number of normal-looking F_2 individuals were tested by growing F_3 progenies from them. If the F_2, though genotypically fired, was phenotypically normal, this will be betrayed by the occurrence of fired individuals in the F_3. Genotypically normal individuals will fail to give fired members in the F_3.

The actual segregations observed were, in the F_2, raised from triple heterozygotes, 161 fired : 276 nor-

mal. Of these normals 60 were tested by growing F_3 progenies from them and of these 5 were found to be genotypically fired and 55 genotypically normal. Are these data in agreement with the above hypothesis ?

If the hypothesis is true we expect 27/64 of the F_2 to be genotypically fired, but of these some portion will look normal. Let us represent this portion by f. Then we expect $(27 - f)/64$ to be phenotypically fired and $(37 + f)/64$ to be phenotypically normal. On testing normals we expect to find $\frac{f}{37 + f}$ of them to be genotypically fired and $\frac{37}{37 + f}$ to be true normals. We thus have two sets of data, each giving an estimate of f in accordance with our hypothesis. If the hypothesis is true we expect that these two estimates of f will be alike. If the data are heterogeneous with respect to the f of the hypothesis, then the hypothesis is wrong. The problem is one of testing heterogeneity by joint estimation.

The joint logarithm likelihood of the two sets of data is :

$$L = 161 \log \frac{27 - f}{64} + 276 \log \frac{37 + f}{64} + 5 \log \frac{f}{37 + f} + 55 \log \frac{37}{37 + f}$$

On summing terms involving like expressions for f and omitting terms independent of f this becomes :

$$L = 161 \log (27 - f) + 216 \log (37 + f) + 5 \log f$$

Differentiating and equating to zero gives

$$\frac{dL}{df} = - \frac{161}{27 - f} + \frac{216}{37 + f} + \frac{5}{f}$$

for the equation of estimation for f.

This reduces to

$$382f^2 + 175f - 4{,}995 = 0$$

and gives $f = 3{\cdot}394254.$

We may now find the expectations for the classes in the F_2 and F_3, using this best fitting value of f. These are set out in Table 28.

TABLE 28

Class	Observed (a)	Expected (mn)	$\chi^2\left(\frac{(a-mn)^2}{mn}\right)$
Fired	161	161·1830	0·0002
Normal	276	275·8170	0·0001
Total	437	437	
Fired	5	5·0417	0·0003
Normal . . .	55	54·9583	0·0001
Total	60	60	0·0007

A χ^2 is calculated for the agreement of the observed values with these expectations. It will have one degree of freedom as one has been lost in fitting f. The probability of obtaining as large or larger deviation is then found to be 0·98 to 0·95. The data are in agreement with the hypothesis.

19.1. FISHER'S SCORING METHOD

It will be recalled from Section 13 that the maximum likelihood equation of estimation is, for a recombination value p,

$$\frac{dL}{dp} = a_1\frac{d\log m_1}{dp} + a_2\frac{d\log m_2}{dp} \ldots + a_t\frac{d\log m_t}{dp} = 0.$$

Now the quantities $\frac{d\log m_1}{dp}$, $\frac{d\log m_2}{dp}$, &c., depend only on the relation m bears to p. For a given type of family m is fixed as an algebraic function of p, so that for any given value of p we can evaluate $\frac{d\log m_1}{dp}$,

&c., once and for all. Then $\frac{d \log m_1}{dp} = \frac{1}{m_1}\frac{dm_1}{dp}$, once evaluated, is the contribution which any individual falling in class 1 will make to the maximum likelihood expression; in other words it is a score by which a_1 can be multiplied to find the contribution of this class to the expression. Summing the products of score and number observed, one such product from each class, will give the total value of the maximum likelihood expression and this total will be 0 if the proper value has been used for p in calculating the scores.

In the same way the total amount of information is $I_p = ni_p = nS\left[\frac{1}{m}\left(\frac{dm}{dp}\right)^2\right]$ so that in any type of family i_p is dependent in value only on the value of p. Then i_p can be evaluated for any value of p and can be multiplied by n to give I_p.

If an incorrect value of p has been used in computing the scores, the value of $\frac{dL}{dp}$ (i.e. D) will not equal 0. We have already seen that I_p is the rate of change of $\frac{dL}{dp}$ on p, so in the region of the true value of p, $\frac{dL}{dp}$ will depart from 0 by a quantity which is the product of the amount of information and the departure, δp, of p from its true value, i.e. $D = I_p\delta p$. Then if D, the deviation from 0, is divided by I_p an estimate is obtained of the correction δp that it is necessary to make to p. This new value of p can then be used in computing new scores and the amount of information and the process repeated. This method is due to Fisher (1946).

Ex. 15.1. The use of Fisher's scoring method in both estimation and testing heterogeneity can be illustrated by the data of Ex. 13. Both of Bateson's sweet pea families were of the F_2 type though one was

actually an F_3 segregating for both genes. Then the expectations, m, scores and contributions of each individual to I_P (all expressed in terms of $P = (1-p)^2$) will be as given in Table 28.1.

TABLE 28.1

Class	m	$\frac{dm}{dP}$	Score $\left[= \frac{1}{m}\frac{dm}{dP}\right]$	$i_P\left[= \frac{1}{m}\left(\frac{dm}{dP}\right)^2\right]$
AB	$\frac{1}{4}(2+P)$	$\frac{1}{4}$	$\frac{1}{2+P}$	$\frac{1}{4(2+P)}$
Ab	$\frac{1}{4}(1-P)$	$-\frac{1}{4}$	$-\frac{1}{1-P}$	$\frac{1}{4(1-P)}$
aB	$\frac{1}{4}(1-P)$	$-\frac{1}{4}$	$-\frac{1}{1-P}$	$\frac{1}{4(1-P)}$
ab	$\frac{1}{4}P$	$\frac{1}{4}$	$\frac{1}{P}$	$\frac{1}{4P}$
Total	1	0	—	$\frac{(1+2P)}{2P(1-P)(2+P)}$

The application of these formulae to Bateson's data is shown in Table 28.2.

TABLE 28.2

Class	Observed	$P = 0{\cdot}80$ Score	$P = 0{\cdot}80$ D	$P = 0{\cdot}843$ Score	$P = 0{\cdot}843$ D
AB	879	0·357143	152·6787	0·351741	0·206547
Ab+aB	96	−5·000000		−6·369427	
ab	255	1·250000		1·186240	
		i_P	I_P	i_P	I_P
Total .	1230	2·901786	3569·1968	3·569208	4390·1258

The F_2 and F_3 data are pooled in the estimation P. $P = 0{\cdot}80$ is taken as the first trial value. The scores are then found from the general formulae of Table 28.1. Thus the score for individuals in class **AB** is

$\frac{1}{2 + P} = \frac{1}{2 + 0{\cdot}8} = 0{\cdot}357143$ and so on. The **Ab** and **aB** classes may be pooled as their scores are the same. The value of i_P is found from the formula $\frac{1 + 2P}{2P(1 - P)(2 + P)}$, or more simply in the present case as one-quarter of the sum of the scores neglecting signs. The latter method is, however, to be used with caution as it is not generally valid.

D is found as the sum of the numbers of individuals in each class multiplied by the score. Thus with $P = 0{\cdot}80$

$$D = (879 \times 0{\cdot}357143) - (96 \times 5{\cdot}000000) + (255 \times 1{\cdot}25000) = 152{\cdot}6787$$

the second terms being negative because the score of this class is negative. We next find

$$I_P = ni_P = 1230 \times 2{\cdot}901786 = 3569{\cdot}1968$$

Then, the correction needed by our trial value of P is calculated as

$$\delta P = \frac{D}{I} = \frac{152{\cdot}678}{3569{\cdot}1968} = 0{\cdot}043$$

Taking $P = 0{\cdot}80 + 0{\cdot}043 = 0{\cdot}843$ as a new trial value we find by a similar calculation (using, of course, scores and i_P evaluated with this new value of P), $D = 0{\cdot}206547$, $I_P = 4390{\cdot}1258$ and $\delta P = 0{\cdot}000047$ which is clearly negligible. We can take $P = 0{\cdot}8430$ as the value correct to four decimal places. Then

$$s_P = \sqrt{V_P} = \sqrt{\frac{1}{I_P}} = \sqrt{\frac{1}{4390{\cdot}1258}} = \sqrt{0{\cdot}000227784} = 0{\cdot}01509$$

The scores obtained using $P = 0{\cdot}8430$ can now be used to test the homogeneity of the two families, F_2 and F_3. The numbers observed in the two families separately are given in Table 23. Using these numbers in conjunction with the scores and i_P appropriate

to $P = 0{\cdot}8430$ we can calculate the contribution of each family to D and to I_P. Thus the contribution of the F_2 to D is $(296 \times 0{\cdot}351741) - (46 \times 6{\cdot}369427) + (85 \times 1{\cdot}186240) = -88{\cdot}0479$ and its contribution to I_P is $427 \times 3{\cdot}569208 = 1524{\cdot}0518$. The contributions of the F_3 are similarly, to D, 88·2545, and to I_P 2866·0740. It will be observed that the sum of the two contributions to D is 0·2066 and the sum of those to I_P is 4390·1258, as was found using the pooled data.

Then the heterogeneity of the families in respect of the linkage values to tested by

$$\chi^2 = \frac{-88{\cdot}0479^2}{1524{\cdot}0518} + \frac{88{\cdot}2545^2}{2866{\cdot}0740} = 7{\cdot}8043$$

for 1 degree of freedom. The families cannot thus be regarded as homogeneous. The slight discrepancy between this value of χ^2 and that of 7·8058 found at the end of Ex. 13 is due to our now using $P = 0{\cdot}8430$ in place of the earlier $P = 0{\cdot}843047$. The difference brought about by the use of the coarser approximation is, of course, so trivial as to involve no possible risk of misjudgement.

CHAPTER VIII

DISTURBED SEGREGATIONS

20. DISTURBED F_2S

SIMPLE formulae have been given in earlier chapters for the detection and estimation of linkage. It is, however, clear that they are only accurate when the single factor segregations are good. The formulae for calculating χ^2 in the detection of linkage are special to given cases ; a change in the segregation of one factor necessitates an entirely new formula. Also the simple application of the method of maximum likelihood will not give an estimate free from error arising out of disturbed single factor ratios. A considerably more complicated use of maximum likelihood could be employed for such cases.

There are, however, methods for the detection and estimation of linkage with disturbed single factor segregations which do give accurate and trustworthy results. These methods do not enjoy some of the advantages of the methods applicable when gene segregation is normal, but are to be preferred to the earlier methods when disturbances are encountered. The methods for the detection of linkage, as given in this chapter, are of wider application than the methods for the estimation of linkage.

We may conveniently divide the treatment of disturbed data into two types, that concerned with single families and that to be used when both coupling and repulsion data are to hand.

Ex. 16. As an example of the analysis of data of

but one type we may take a case of linkage between a gene and a second one which is a member of a pair of complementary factors. Then the character controlled by the second linked gene will not segregate in a 3 : 1 ratio but will give a 9 : 7 in the F_2. As we know what the true segregation is we can compare our approximate results, obtained by treating the 9 : 7 as a disturbed 3 : 1, with the true values obtained by using the 9 : 7 expectation. We may utilize data from Jenkins (1927) concerning the segregation for green and yellow plant colour and purple and white aleurone colour in maize. The former is controlled by a single factor and the latter by a pair of complementary factors. In one family segregating for all the factors the segregation observed was :

Green Purple	Yellow Purple	Green White	Yellow White
127	19	67	44

We desire to know if there is linkage between the genes, assuming that the purple-white segregation is due to a single factor disturbed by unknown causes, and if linkage is found, what the recombination value is.

We first note that the segregation for green : yellow is 194 : 63 and that this is a good 3 : 1 (as tested by χ^2). On the other hand, the segregation for purple : white cannot be considered as agreeing with an expectation of 3 : 1 when tested by χ^2. Then we cannot use the ordinary method for calculating χ^2 to test for linkage. We may, however, calculate the linkage χ^2 in a different manner. The data are set out for this purpose in a 2 × 2 table (contingency table) thus :

TABLE 29

127	19	146
67	44	111
194	63	257

The marginal totals give the single factor segregations of the two genes. If there is no linkage between the genes we may expect that the frequency of any one class observed in the experiment is proportional to the corresponding marginal frequencies. Thus the green purple class would be expected to occur in $\frac{146}{257} \times \frac{194}{257}$ of cases. As there are 257 in the family we then expect 110·2101 in this class. In this way we could calculate the expectation for the other classes and then find χ^2 from the formula $S\left(\frac{a^2}{nm}\right) - n$. This χ^2 would be a test of departure from independence of segregation of the genes, i.e. of linkage.

We can, however, calculate the χ^2 in a much simpler manner, viz. by using the formula

$$\chi^2 = \frac{(a_1a_4 - a_2a_3)^2n}{(a_1 + a_2)(a_3 + a_4)(a_2 + a_4)(a_1 + a_3)}$$

where a_1, a_2, &c. and n have the same meaning as in earlier chapters. Applying this formula we find:

$$\chi^2 = \frac{(127 \times 44 - 19 \times 67)^2 \times 257}{146 \times 111 \times 63 \times 194} = 24{\cdot}159$$

This corresponds to one degree of freedom as, of the original three, two have been taken up in using the observed marginal totals as the best fitting segregations for the single factors. The significance of this χ^2 is beyond question.

As a check on the method we may calculate the χ^2 for linkage, utilizing the knowledge that the purple-white segregation is one of 9 : 7. The formula for the χ^2 detecting linkage in such a family, arrived at by the methods of Chapter VI, is

$$\chi^2 = \frac{(7a_1 - 21a_2 - 9a_3 + 27a_4)^2}{189n}$$

and in this case $\chi^2 = 23{\cdot}792$. The difference **between**

the linkage χ^2s as calculated in these two ways is negligible. On the other hand, if we calculate χ^2 from the formula $\frac{(a_1 - 3a_2 - 3a_3 + 9a_4)^2}{9n}$, which is correct when both factors are giving a 3 : 1 ratio, we find that it is 30·361, a misleadingly high value. In the present case the overestimation of the significance of the evidence for linkage is not serious, but it is easy to see that it could be highly misleading in cases where the evidence is more doubtful.

We may now consider the estimation of linkage. In place of the method of maximum likelihood, which cannot be used when the single factor segregation is poor, we may employ the product method. This method will in certain cases give an absolutely accurate estimate of linkage and in any case will reduce the error arising from the poor segregation as compared with the usual method of maximum likelihood formula for undisturbed F_2s.

The product formula puts $\frac{a_1a_4}{a_2a_3} = \frac{2P + P^2}{1 - 2P + P^2}$ for the usual two factor F_2. Applying it to the present case we find

$$\frac{5588}{1273} = \frac{2P + P^2}{1 - 2P + P^2}$$

or $$4315P^2 - 13722P + 5588 = 0$$

Then $$P = 0{\cdot}479542$$

and $$p = 1 - \sqrt{P} = 0{\cdot}3075$$

This value of p is rather far from the true value of p as calculated when the 9 : 7 is treated as such and not as a bad 3 : 1. It is, however, considerably nearer to the correct value than is the estimate reached if the data are treated as a simple two factor F_2 by maximum likelihood (35 per cent). Thus although the error is not completely removed it is reduced.

This example is, however, not one which shows the

product formula at its best. Let us take an artificial example where the ratios are disturbed by poor manifestation of the recessive. We may then compare the estimates of the recombination fraction with the value used as the basis for the construction of the example.

If we take an F_2 for two linked factors with 25 per cent recombination in repulsion we expect a segregation of 2·0625 **AB** : 0·9375 **Ab** : 0·9375 **aB** : 0·0625 **ab**. Let us further consider that the recessive **bb** failed to manifest itself in 40 per cent of cases. The 40 per cent of the **Ab** and **ab** classes will be classified as **AB** and **aB** respectively. Then we expect to find an observable segregation of 2·4375 **AB** : 0·5625 **Ab** : 0·9625 **aB** : 0·0375 **ab**. Calculation of the recombination fraction from these figures by the simple maximum likelihood F_2 formula (Chapter V) gives $P = 0{\cdot}077972$ $p = \sqrt{P} = 0{\cdot}2792$. This is a deviation of 2·92 per cent from the real value of 25 per cent.

If, on the other hand, we calculate P, and from it p, by the product method we obtain a much better result. The equation of estimation (*see* above) is

$$\frac{0{\cdot}09140625}{0{\cdot}54140625} = \frac{2P + P^2}{1 - 2P + P^2}$$

and we find $$P = 0{\cdot}70457$$
$$p = 0{\cdot}2654$$

The deviation is now but half of that shown by the maximum likelihood estimate. This well illustrates the effect of the product method in minimizing such errors. If the disturbance had been due solely to poor viability of the **bb** classes, the product formula would have given an absolutely correct estimate.

It should be remembered that when we say that the method of maximum likelihood gives an estimate showing considerable error we do not mean that the *correct* application of maximum likelihood will give a wrong estimate. If, in setting up the expected values, we can, by utilization of some hypothesis as

to their nature, allow for the disturbances in the segregation a correct estimate would be obtained. It is in cases where such a knowledge of the cause of disturbance is not possessed that the product formula is of more value.

According to the nature and degree of disturbance, the product method is either fully, or nearly fully, efficient for the estimation of linkage. Its full efficiency with undisturbed F_2s is demonstrated at the end of this chapter. Its slight loss of efficiency in backcrosses and F_2s with disturbed viability has been examined by Bailey (1949).

21. THE EXACT TREATMENT OF BACKCROSS DATA

In the detection and estimation of the recombination fraction it has been supposed above that the disturbance in the segregation of one factor has not affected the segregation of the other. This may not be true in all cases, though in this example the good segregation of the first factor suggests that this assumption is not incorrect. In general, failure of manifestation of one character, due for example to incomplete penetrance, will give results justifying the use of this method. On the other hand, reduced viability of one factor will affect the segregation of anything linked to it and so the method may not be completely suitable.

Where both factors are showing disturbed segregation this method must be used with considerable caution. If one class is very short in numbers, so causing the disturbed segregations of both factors, it cannot be treated in this way.

The second approach to the detection and estimation of linkage where gene ratios are disturbed, as developed by Fisher (1949*b*), is beyond these criticisms and is frequently applicable. It demands, however, the joint use of coupling and repulsion data. This second method has been developed for the backcross only. It is not clear that it can be used uncritically for F_2

data, as they do not share with the backcross the characteristic of each phenotypic class comprising but one genotypic class. In this case the corrections for viability may not be absolutely correct. It would, however, certainly give an exact test for the detection and a first approximation to the correct treatment in estimation of disturbed F_2 data.

Ex. 17. Nabours (*et. al.* 1933 and unpublished data supplied to R. A. Fisher) finds the following segregations on backcrossing Grouse Locusts (*Acridium arenosum*), heterozygous for the two factors **W** and **My**, to the double recessive.

TABLE 30

	w my	**W my**	**w My**	**W My**	Total
Repulsion .	30	70	2	24	126
Coupling. .	519	119	12	349	999

It will be seen that in both cases the **w My** class is very small as compared with any other. Neither **W,w** nor **My,my** is giving a good 1 : 1 segregation such as would be expected from a backcross. This is largely attributable to the shortage of **w My** animals, though it is to some extent aggravated by a small shortage of **W My** locusts. It is, however, clear that the disturbance of each factor is due to its interaction in viability with the other factor. Hence the methods of the previous example cannot be used. We must proceed by comparison of the two families.

First let us add the **w my** and **W My** animals together, calling the sum A_1. Similarly the **W my** and **w My** animals are summed to give A_2. This is done for the coupling and repulsion data separately.

Then in repulsion the A_1 class is one of recombination individuals, and A_2 comprises the parental combinations. In the coupling data the reverse is, of course, the case. These classes should be potentially equal in the absence of linkage, though viability disturbance may reduce one or other of them. No matter what the cause of the disturbance it should

affect the A_1 class in coupling and A_2 class of repulsion equally, as they comprise the same two genotypes in potentially equal numbers. Then in the absence of linkage the ratio of A_1 to A_2 should be the same in both sets of data irrespective of *any* viability disturbance. This expected similarity provides the basis for the detection of linkage. The data are set out as in Table 31.

TABLE 31

	A_2	A_1	
Repulsion	72	54	126
Coupling	131	868	999
	203	922	1125

The marginal totals are found and then a χ^2 testing the hypothesis that the observed four classes are proportional to the marginal totals, i.e. that the $A_1 - A_2$ subdivision is independent of the coupling-repulsion subdivision (as it will be in the absence of linkage) may be calculated. It is done by the same formula used for the table in the last example. We find

$$\chi^2 = \frac{(868 \times 72 - 131 \times 54)^2 1125}{126 \times 999 \times 203 \times 922} \doteq 146{\cdot}674$$

for one degree of freedom. This is clearly of very great significance and there can be no doubt of the existence of linkage.

We next ask the value of the recombination fraction between the genes.

Now, if we imagine two hypothetical values α_1, α_2 of the type of A_1 and A_2 but undisturbed by viability differences it is clear that

$$\frac{p}{1-p} = \frac{\alpha_1}{\alpha_2} \qquad \text{for repulsion}$$

and

$$\frac{p}{1-p} = \frac{\alpha_2}{\alpha_1} \qquad \text{for coupling}$$

But the ratio $\dfrac{A_1}{A_2}$ departs from the ratio $\dfrac{\alpha_1}{\alpha_2}$ because of

viability troubles. The departure from this perfect ratio, caused by the inviability of some genotypes, is the same in the coupling and repulsion data. We can thus write :

$$\frac{A_1}{A_2} = \frac{v\alpha_1}{\alpha_2}$$

Then for repulsion $\dfrac{p}{1-p} = \dfrac{A_{R1}}{vA_{R2}}$

and for coupling $\dfrac{p}{1-p} = \dfrac{vA_{C2}}{A_{C1}}$

Hence simple multiplication gives

$$\left(\frac{p}{1-p}\right)^2 = \frac{A_{R1}A_{C2}}{A_{R2}A_{C1}}$$

where A_{R1} is A_1 in the case of repulsion, &c. This equation provides a method of estimating p independently of the viability disturbance v. It may be noted here that the equation

$$\frac{A_{R1}A_{C1}}{A_{R2}A_{C2}} = v^2$$

itself also derived by the above considerations, provides an estimate of v independently of p. The method may be used for either purpose.

Applying this method to the estimation of p from Nabour's data we find

$$\frac{p^2}{(1-p)^2} = \frac{54 \times 131}{72 \times 868} = 0{\cdot}113191$$

Then $\dfrac{p}{1-p} = 0{\cdot}33644$

and $p = \dfrac{0{\cdot}33644}{1{\cdot}33644} = 0{\cdot}2517$ or 25·17 per cent.

Our next concern is to find an estimate of the variance of the statistic p. There exists a simple formula for this purpose (cf. Fisher, 1936*b*). The

derivation of this formula is given in the last section of this chapter. The formula itself is

$$V_p = \frac{p^2(1-p)^2}{h}$$

where h is the harmonic mean of A_{R1}, A_{R2}, A_{C1}, and A_{C2}, i.e.

$$\frac{1}{h} = \frac{1}{4}\left(\frac{1}{A_{R1}} + \frac{1}{A_{R2}} + \frac{1}{A_{C1}} + \frac{1}{A_{C2}}\right)$$

In the present case we find $h = 97{\cdot}102$ and $p = 0{\cdot}2517$.

Then $\quad V_p = \dfrac{(0{\cdot}2517)^2(0{\cdot}7483)^2}{97{\cdot}102} = 0{\cdot}0003653$

and $\quad s_p = \sqrt{V_p} = 0{\cdot}01911$

22. THE CALCULATION OF VARIANCE FORMULAE

In the above examples certain assertions were made about the formulae for the variances of the statistics used. The methods by which such variances are obtained may be illustrated by the derivation of the two formulae used in the last two examples. First consider the question of the variance of the statistic P calculated from an F_2 by the product formula method (Fisher 1936*a*). This formula puts

$$\frac{a_1a_4}{a_2a_3} = \frac{2P+P^2}{1-2P+P^2}$$

Then, taking logarithms,

$$\begin{aligned} F &= \log a_1 + \log a_4 - \log a_2 - \log a_3 \\ &= \log(2+P) + \log P - 2\log(1-P) \end{aligned}$$

The variance of F may be found from the general formula

$$\frac{1}{n}V_F = S\left[m\left(\frac{dF}{da_1}\right)^2\right] - \left(\frac{dF}{dn}\right)^2 \text{ where } n = S(a)$$

Now $\quad \dfrac{dF}{da_1} = \dfrac{1}{a_1},\ \dfrac{dF}{da_2} = -\dfrac{1}{a_2},\ \dfrac{dF}{da_3} = -\dfrac{1}{a_3},$

$$\frac{dF}{da_4} = \frac{1}{a_4},\ \frac{dF}{dn} = 0$$

Then substituting expectation for a_1 &c., we find

$$\frac{1}{n}V_F = \frac{\frac{1}{4}(2+P)}{\frac{n^2}{16}(2+P)^2} + \frac{\frac{1}{4}P}{\frac{n^2}{16}P^2} + \frac{\frac{2}{4}(1-P)}{\frac{n^2}{16}(1-P)^2}$$

or

$$\frac{n}{4}V_F = \frac{1}{2+P} + \frac{1}{P} + \frac{2}{1-P} = \frac{2(1+2P)}{P(2+P)(1-P)}$$

To obtain V_P from V_F we make use of the general formula $V_P = V_F \div \left(\frac{dF}{dP}\right)^2$ already used in Chapter V,

$$\frac{dF}{dP} = \frac{1}{2+P} + \frac{1}{P} + \frac{2}{1-P} = \frac{2(1+2P)}{P(2+P)(1-P)}$$

Then $$\frac{n}{4}V_P = \frac{P(2+P)(1-P)}{2(1+2P)}$$

and $$V_P = \frac{2P(2+P)(1-P)}{n(1+2P)}$$

This is also the formula for the variance of the maximum likelihood statistic, and so the product formula gives a fully efficient estimate of P in this case.

The second example (Fisher, 1949*b*) discussed in this chapter utilized the estimation equation

$$\frac{p^2}{(1-p)^2} = \frac{A_{R1}A_{C2}}{A_{R2}A_{C1}} = \frac{\alpha_{R_1}\alpha_{C_2}}{\alpha_{R_2}\alpha_{C_1}}$$

What is the variance of p?

Using the same notation as in the example let us consider the simple case $\frac{p}{q} = \frac{\alpha_1}{\alpha_2}$ where $q = 1 - p$

$$\log\frac{p}{q} = \log p - \log q$$

and $$\frac{d}{dp}\left(\log\frac{p}{q}\right) = \frac{1}{p} + \frac{1}{q} = \frac{1}{pq}$$

I_x *the amount of information about* x is the inverse of the variance of x.

Then

$$I_{\log \frac{p}{q}} = I_p\left(\frac{dp}{d_{\log \frac{p}{q}}}\right)^2 = I_p p^2 q^2 = \frac{np^2q^2}{pq} = npq$$

or putting

$$p = \frac{\alpha_1}{n} \text{ and } q = \frac{\alpha_2}{n},\ I_{\log \frac{p}{q}} = \frac{\alpha_1\alpha_2}{n}$$

or

$$V_{\log \frac{p}{q}} = \frac{1}{\alpha_1} + \frac{1}{\alpha_2}$$

But we actually estimate $\frac{p}{q}$ as the geometric mean of $\frac{A_{R_1}}{A_{R_2}}$ and $\frac{A_{C2}}{A_{C1}}$ (i.e. as the geometric mean of $\frac{\alpha_{R1}}{\alpha_{R2}}$ and $\frac{\alpha_{C2}}{\alpha_{C1}}$)

Remembering that if

$$V_{\log \frac{p}{q}} = \frac{1}{A_1} + \frac{1}{A_2}$$

then

$$V_{\frac{1}{2}\log \frac{p}{q}} = \frac{1}{4}\left(\frac{1}{A_1} + \frac{1}{A_2}\right)$$

the variance of $V_{\log \frac{p}{q}}$ as estimated in this manner is found by

$$V_{\log \frac{p}{q}} = \frac{1}{4}\left(\frac{1}{A_{R1}} + \frac{1}{A_{R2}} + \frac{1}{A_{C1}} + \frac{1}{A_{C2}}\right)$$

$$= \frac{1}{h}$$

where h is the harmonic mean, because the variance

of a sum is the sum of the variances of the two components. Now

$$V_p = V_{\log \frac{p}{q}} \div \left(\frac{d \log \frac{p}{q}}{dp} \right)^2$$

$$= \frac{1}{h} \div \left(\frac{1}{pq} \right)^2 = \frac{p^2 q^2}{h}$$

This is the formula utilized in the example.

CHAPTER IX

HUMAN GENETICS (I)

23. HUMAN DATA

THERE are two characteristics of human genetical data that make its statistical reduction different from, and rather more complex than, that applied to normal genetical results. These are respectively the small size of the families produced by any mating and the incomplete information available about the type of mating. These difficulties are overcome by the development of a correspondingly more elaborate statistical technique. It should be noticed, however, that although the statistical treatment is superficially different from that described for non-human material, it is characterized by certain fundamental similarities to the methods developed in the previous chapters. The following description of the statistical methods applicable to human material is not intended to be complete, but, it is hoped, will serve as an indication of the general line of approach to these special problems. More detailed analyses will be found in the various articles cited in the text.

The two main statistical problems, viz. single factor segregation and linkage, will be considered separately and in that order.

24. SINGLE FACTOR SEGREGATIONS

The two difficulties noted above as characteristics of human data are encountered immediately a consideration of single factor segregation is undertaken.

In the first place the smallness of the families invalidates the use of the χ^2 test of deviation from the expected segregation. Where expectation in any class is less than 5, the χ^2 calculated from that family departs seriously from the tabulated large sample χ^2 distribution. Hence it cannot be used as a test of significance in such cases. This difficulty can, of course, be overcome by suitable lumping in order to test the significance of deviations, but in testing heterogeneity the difficulty is felt with full force. Until an easily applicable generalized χ^2 is available, this test is ruled out of general use.

The second characteristic, that of incomplete information about the type of mating, is perhaps even more troublesome but can be overcome. Consider the case of a rare recessive character in a population. With random mating the frequencies of the genotypes **AA**, **Aa** and **aa** will be $(1-p)^2 : 2p(1-p) : p^2$ where p is the proportion of gametes carrying the recessive allelomorph. When p is small p^2 is very small (e.g. $p = 0{\cdot}01$ gives **AA** 0·9801 : **Aa** 0·0198 : **aa** 0·0001). Then matings involving **aa** individuals will be extremely rare. Segregating matings in which one parent is **aa** will be even rarer. Hence we may assume, in order to remove this uncertainty, that all families showing segregation for the recessive character will be from matings of two heterozygotes, **Aa** × **Aa**. The error involved in this assumption will be small when p is small, but if p is large will become important. Fortunately many of the hereditary human characters are rare conditions.

The small size of the families has, in addition to its effect in invalidating χ^2 as a test of significance, another troublesome effect. All matings of the type **Aa** × **Aa** will not give recessives among the progeny. Some will give all normals. These cannot then be distinguished from matings of which one parent was **AA**. Now such families will be lost to the records. We must, then, have some procedure based solely on

families with one or more recessive **aa** children. Such families may be found by one of two procedures, or a mixture of both, viz. searching whole communities or sections of communities for affected families or by finding recessive individuals and following up their pedigrees. In the former case of 'complete ascertainment', all segregating families, no matter how many recessives they may contain, will be included once. In the latter case of 'ascertainment through affected individuals', the chance of finding and recording the family is clearly proportional to the number of recessives in it. The investigator is twice as likely to meet one or other or both of two recessives as to meet a single individual.

Various methods of handling data of these types have been suggested. Some, like the proband method, are solely of value in the case of complete ascertainment. In other cases they may give misleading results. Now complete ascertainment is difficult and rarely achieved. Consequently such methods are of little general value.

Of all the methods the sib treatment is the most generally applicable. It takes into account the chance of ascertainment where this is through affected individuals, and is also applicable to completely ascertained data by a simple extension of the argument. There is a small loss of efficiency as compared with the proband method in this latter case, but much of the lost information can be recovered, if desired, by a more complicated analysis (*see* Fisher, 1934).

25. THE SIB METHOD

The logic of the sib method is simple. The chance of any sib of an affected individual being itself affected is independent of the affected nature of the first sib. Then by adding up the frequencies of the normal and affected individuals among the sibs of affected individuals a good estimate of the proportion of recessives emerging in segregating families will be

obtained. It is essential to the avoidance of biased results, when using this method, that any family be used as often as it is ascertained, because the method is based on sampling the *sibs of affected members* of the population and not on sampling the families with affected members. If the family is found by virtue of one affected sib it is used once. If it is ascertained through each of five affected members it must be entered in the records five times. This method then allows for or even demands that the frequency of ascertainment be proportional to the number of affected members it contains. Complete ascertainment is included in this scheme by considering such families as though ascertained through each one of their affected members.

The working of the method may be simply demonstrated, using families of three children. Such families, produced by the cross **Aa** × **Aa**, will contain 0, 1, 2 or 3 **aa** children with the frequencies given by the expansion of $(\frac{3}{4} + \frac{1}{4})^3$. These are set out in the second column of Table 32.

TABLE 32

Affected sibs in family of three	Frequency	Sibship scores Normal	Sibship scores Affected
0	$\frac{27}{64}$		
1	$\frac{27}{64}$	$\frac{54r}{64}$	0
2	$\frac{9}{64}$	$\frac{18r}{64}$	$\frac{18r}{64}$
3	$\frac{1}{64}$	0	$\frac{6r}{64}$
Total . .	1	$\frac{72r}{64}$	$\frac{24r}{64}$

The first type of family with 0 affected children

cannot be ascertained as distinct from the progeny of other crosses, and so is omitted from further considerations. The next type of family will be ascertained via its single affected sib. Let us suppose that the chance of encountering any affected individual is r. Then such families will be found in r of cases and, as both the sibs of the affected child are normal, it will contribute $\frac{27}{64} \times r \times 2$, i.e. $\frac{54r}{64}$ to the column headed 'Normal' in that section of the table given to the sibship scores. There are no affected sibs and consequently no contribution to the 'Affected' score.

Families containing two affected individuals will have a chance $2r$ of entering the records. Each family contains one normal and one affected sib in addition to the one through which ascertainment was made. Hence the contribution to the 'Normal' and 'Affected' columns are both $\frac{9}{64} \times 2r \times 1$, i.e. $\frac{18r}{64}$.

Similarly families with three affected children will be found in $3r$ of cases and will contribute nothing to the 'Normal' column, but $\frac{1}{64} \times 3r \times 2$, i.e. $\frac{6r}{64}$ to the 'Affected' column.

The sums of the 'Normal' and 'Affected' columns are then found to be $\frac{72r}{64}$ and $\frac{24r}{64}$. This is the typical 3 : 1 ratio of a single factor F_2.

The method can be demonstrated in a similar manner for the general case of a family of size n and a segregation of $x : y$.

It is clear that with a number of families this method can give a test of deviation from the expected ratio. The variation in the summed scores will depend on variations in family size, frequencies of possible types in any size of family, and on the frequency of ascertainment. All these will contribute

to the standard error of the score and must be employed in the test of significance.

The formula for the variance of y, the proportion of recessives for any size of family, is

$$V_y = \frac{1}{n'} \frac{y(1-y)}{s-1}(1 + y' + 2yy')$$

where n' is the total number of affected individuals ascertained, s is the family size and y' a measure of the completeness of ascertainment. y' is calculated from the formula

$$y' = \frac{S\{t(t-1)n_{at}\}}{S\{t(a-1)n_{at}\}}$$

where t is the number of ascertainments, a the number affected and n_{at} the number of cases in class at, i.e. the class with a affected and t ascertained (Fisher 1934)

Ex. 18. To make this procedure clear we may take the question of the proportion of albinos in families segregating for this character. Is this proportion the 0·25 that would be expected if albinism is a simple autosomal recessive ?

The following forty-seven families of five, six, and seven children were found by Pearson, Nettleship and Usher (1913) to be segregating various numbers of albinos as shown in Table 33.

TABLE 33

	Size of Family		
No. of Albinos	5	6	7
1	7	4	4
2	6	6	4
3	4	3	5
4	1	1	1
Total . .	18	14	15

Consider first the families of five children. We may suppose complete ascertainment in this case,

in the absence of contradictory evidence. Then the data may be set down in the form of Table 34.

TABLE 34

No. of Albinos	No. of Families	Sibship scores Normal	Sibship scores Affected
1	7	28	0
2	6	36	12
3	4	24	24
4	1	4	12
Total	18	92	48

The families with 1 albino, 7 in number, will each contain 4 normal sibs of the albino and so will contribute 4 × 7 to the ' Normal ' column, and 0 to the ' Affected ' column (r is assumed to be 1).

The six families with 2 albinos have, for each albinotic child, 3 normal and 1 albinotic sibs. The families must be counted twice as we suppose them to have been found through each of the 2 albinos. Their contributions to the normal and affected scores will be 3 × 6 × 2 and 1 × 6 × 2 respectively.

The entries for the remaining families with 3 or 4 albinos may be calculated in a similar manner. On summing these columns we find 80 normal and 36 affected sibs. Then $y = \frac{48}{92 + 48} = 0{\cdot}342857$.

Now y', the measure of completeness of ascertainment of affected children, is 1 as we have assumed complete ascertainment. The number of ascertainments, n', is here the total of affected children, i.e. 35, and $s - 1$ is 4. As we are testing agreement with the hypothesis of $y = 0{\cdot}25$ we must use this value of y in calculating the variance (cf. the standard error of 3 : 1 ratios in Chapter II).

$$\text{Then } V_y = \frac{1}{35}\,\frac{\frac{3}{4} \times \frac{1}{4}}{4}(1 + 1 + \tfrac{1}{2}) = 0{\cdot}00334821$$

$$\text{and } \quad \sigma_y = \sqrt{V_y} = 0{\cdot}05787$$

Thus the deviation of the observed y from its expected 0·25 is 0·342857 − 0·25 or 0·092857 ± 0·057873. Such a deviation is not significant. The families of 5 agree with the single factor hypothesis.

The values of y and V_y for families of 6 and 7 are arrived at in the same way. They are set out in Table 35.

TABLE 35

Family Size	y	V_y	σ_y	I_y	χ^2
5 .	0·342857	0·00334821	0·057873	298·7	2·575
6 .	0·289655	0·00323276	0·056857	309·3	0·358
7 .	0·274725	0·00252016	0·050201	396·8	0·243
Mean	0·299575	0·0009952	0·03155	1004·8	2·470

We have now three independent estimates of y, each with its own variance. A compound estimate of y may be obtained by finding the weighted mean of the three separate estimates, the weights being the amounts of information (i.e. reciprocals of variances) concerning the various estimates.

Then $$\bar{y} = \frac{S[I_{y_1}y_1]}{S[I_{y_1}]} \text{ where } I_{y_1} = \frac{1}{V_{y_1}}$$

and $$V_{\bar{y}} = \frac{1}{S[I_{y_1}]}$$

Using these formulae we find from Table 35

$$\bar{y} = 0{\cdot}299575 \quad V_{\bar{y}} = 0{\cdot}0009952 \quad \sigma_{\bar{y}} = 0{\cdot}03155$$

Then the deviation of y from the expected 0·25 is

0·04958 ± 0·03155

This is not significant and so the data all agree with the Mendelian expectation.

As we have at hand these three independently estimated values of y we may perform a simple test of heterogeneity. It must clearly be based on the differences between the estimates of y afforded by families of different sizes.

It will be remembered that, in Chapter II, we noted that the χ^2 testing the deviation of a ratio from its expectation is given by

$$\chi^2 = \frac{(y - \nu)^2}{V_y}$$

where ν is the expected value of y.

Applying this to the data from families of 5 individuals we find

$$\chi^2 = \frac{(0{\cdot}34286 - 0{\cdot}25)^2}{0{\cdot}0033482} = 2{\cdot}575$$

This is entered in the last column of Table 35. The χ^2 for families of 6 and 7 are found similarly. We also find and enter the χ^2 from the weighted mean value of y. This last measures the joint deviation from the hypothesis.

Then by adding the three χ^2s from the three family sizes and subtracting the χ^2 of the joint estimate we obtain an analysis similar to that for ordinary genetical segregations. This analysis is :

TABLE 36

	χ^2	D.f.	P.
Deviation	2·470	1	0·20 – 0·10
Heterogeneity	0·706	2	0·80 – 0·70
Total	3·176	3	

As no value of χ^2 is significant it can be said that the data on albinotic children agree both with one another and with the Mendelian expectation of $\frac{1}{4}$ affected in segregating families.

The similarity of this method with those of Chapter II is obvious. Both tests involve the finding of a quantity and its variance. This is the material of the test of significance of the departure from expectation either by the use of the standard error or χ^2. The difference of the two treatments, non-human and human, lies in the necessity for finding a new suit-

able quantity y, of known expectation, in the latter case.

It may be mentioned that the sibship method is also valuable with other types of data. One other use, to which it has already been put, is the estimation of ovule sterility in *Pisum*. The pods correspond to the families in the above example. Fertile and sterile ovules may be recorded at harvest with the important limitation that pods having no fertile ovules are lost as they fail to develop and drop from the plant. The problem is a replica of that worked out above and has been successfully treated by the sib method.

CHAPTER X

HUMAN GENETICS (II)

26. LINKAGE

THE question of linkage detection and estimation from human pedigrees is complicated by the same two difficulties, incomplete knowledge of the mating and small families, as is the consideration of single factor segregation.

Where there is knowledge of three or more generations in the pedigree it is often possible to decide on the nature of the cross, and by lumping families from matings of the same type, to deal with the data by the methods adapted to the more usual types of genetical material.

Ex. 19. For example, Haldane (1936) gives certain families segregating for retinitis pigmentosa, an eye defect. This anomaly is due to one of the so-called dominant genes, i.e. the heterozygote is the type usually distinguished from the normal. The problem at issue was that of whether the gene for retinitis pigmentosa was incompletely linked to sex or not.

The informative families are those from the mating of affected men and normal women. The male parent is then heterozygous for both sex and the eye defect. It is, however, also necessary to know the phase of the linkage, whether coupling or repulsion. With a dominant gene such information is usually easy to obtain. These retinitis pigmentosa pedigrees give the required information in telling whether the man in question received the defect from his father

or his mother. If he received it from his father it will have been transmitted to him with his Y chromosome, and if from his mother with his X chromosome. In the former case, which we may term coupling, we expect an excess of normal females and affected males among the man's progeny, on the hypothesis of incomplete sex-linkage. Similarly, if the man in question received the gene with his X chromosome from his mother we may expect his progeny to contain excess of normal sons and affected daughters. This may be termed the repulsion case.

The results quoted by Haldane (l.c.) are from matings of normal women and retinitis pigmentosa men, and may be summarized as Table 37.

TABLE 37

	Affected		Normal		
	Males	Females	Males	Females	Total
Coupling . .	50	30	27	26	133
Repulsion . .	30	31	57	37	155

The coupling data were from thirty-two families and the repulsion data from thirty-three families. The results are suggestive of sex linkage, but the single factor ratios are somewhat disturbed. As they are double backcross families (**XY Rr** $\times$ **XXrr**) we may use the method of Ex. 17 in testing for and estimating linkage. We then find :

TABLE 38

	A_1	A_2	
Coupling	57	76	133
Repulsion	88	67	155
Total	145	143	288

$$\chi^2 = 30{\cdot}167 \text{ for one degree of freedom}$$

There can be no question of the significance of the evidence for linkage.

Proceeding to the estimation of p we find

$$\frac{p^2}{(1-p)^2} = \frac{57 \times 67}{88 \times 76}, \text{ i.e. } p = 0{\cdot}4304$$

and $$s_p = \sqrt{\frac{p^2(1-p)^2}{h}} = 0{\cdot}0293$$

Haldane further considers the possibility of there being other and autosomal genes producing this eye defect, but we need not deal with this question in detail. It is sufficient to note that such genes will not affect the test of significance but may affect the estimate of p.

27. THE u STATISTICS

In addition to these families for retinitis pigmentosa Haldane also quotes families segregating for recessive characters which it is desired to test for incomplete sex linkage. In these cases there is seldom any clue as to whether the heterozygous male parent received the gene for the defect from his father or his mother, i.e. whether he is doubly heterozygous in coupling or repulsion. Further, the small size of the human family seldom allows of the question being decided from the progeny of such ambiguous males. It is thus impossible to apply the same technique as to retinitis pigmentosa. Various other methods have been suggested for the solution of this problem, but the most generally efficient method is that of Fisher (1935*a*, 1935*b*, 1936*c*). Certain quantities, denoted by u, are calculated and used as the basis of the decision. The precise formula for u varies with the type of family. For the double backcross (**AaBb** × **aabb**) we take

$$u_{11} = (a_1 - a_2 - a_3 + a_4)^2 - (a_1 + a_2 + a_3 + a_4)$$

for the single backcross (**AaBb** × **Aabb**)

$$u_{31} = (a_1 - 3a_2 - a_3 + 3a_4)^2 - (a_1 + 9a_2 + a_3 + 9a_4)$$

and for the F_2 (**AaBb** × **AaBb**)

$$u_{33} = (a_1 - 3a_2 - 3a_3 + 9a_4)^2 - (a_1 + 9a_2 + 9a_3 + 81a_4)$$

The affinities of these formulae with those for the linkage χ^2 of similar families are several. In fact $u_{11} = n(\chi^2 - 1)$. The chief point to note about these u statistics is that, like χ^2, they test equally well deviations from expectation whether in the direction indicating coupling or in the reverse way. In fact, it can be shown that u is a measure of $p(1 - p)$. Since coupling families showing p recombination may be written as repulsion families showing $1 - p$ recombination, it is then clear that the value of u is independent of the phase of the linkage. We may take Haldane's data on the segregation of the recessive achromatopsia to illustrate the use of the u statistics.

Ex. 20. The twenty-eight families given by Haldane are set out in Table 39. The first four columns give the number of normal and affected males and females in the family and the fifth and sixth give the family size and the number of affected individuals. The seventh column gives the value of

$$u_{31}=(a_1-3a_2-a_3+3a_4)^2-(a_1+9a_2+a_3+9a_4)$$

for each family. We note that u_{31} is the correct statistic to use as the cross is of normal male by normal female, i.e. **XY Aa** $\times$ **XXAa**, which is a single backcross. The 28 values of u_{31} are summed and the sum is a measure of $1 - 4x$ where x is an estimate of $p(1 - p)$.

In order to obtain the actual value of $1 - 4x$ it is necessary to divide $S(u_{31})$ by a divisor, $S(k)$, depending on the method of ascertainment. The quantity k is calculated for each family and its sum is the divisor. In the case of complete ascertainment, we take

$$k_c = s(s - 1)\frac{4^s - 3^{s-2}}{4^s - 3^s}$$

which is tabulated by Fisher (1935*a*).

TABLE 39

ACHROMATOPSIA (Haldane's data)

Males		Females							
Normal	Affected	Normal	Affected	s	s_2	u_{31}	k_c	k_s	k_l
1	3	4	0	8	3	112	61·538	84	86·$\dot{2}$
1	0	2	2	5	2	4	25·531	36	30·$\dot{6}$
2	1	3	1	7	2	− 22	47·751	66	40·$\dot{2}$
2	2	3	0	7	2	26	47·751	66	40·$\dot{2}$
1	0	1	2	4	2	16	16·937	24	26·$\dot{2}$
2	2	1	0	5	2	4	25·531	36	30·$\dot{6}$
0	1	1	1	3	2	− 18	9·892	14	22·0
3	2	1	1	7	3	− 30	47·751	66	79·$\dot{3}$
2	1	0	0	3	1	− 10	9·892	14	4·$\dot{2}$
0	6	2	0	8	6	344	61·538	84	294·$\dot{2}$
2	2	1	4	9	6	− 8	77·196	104	306·$\dot{6}$
1	2	0	1	4	3	− 24	16·937	24	60·0
1	0	3	1	5	1	− 12	25·531	36	9·$\dot{3}$
1	3	0	0	4	3	36	16·937	24	60·0
0	0	0	2	2	2	18	4·286	6	18·0
2	3	1	0	6	3	34	35·574	50	72·$\dot{6}$
0	1	0	1	2	2	− 18	4·286	6	18·0
2	1	0	4	7	5	74	47·751	66	200·$\dot{2}$
0	0	0	2	2	2	18	4·286	6	18·0
3	0	2	1	6	1	2	35·574	50	12·$\dot{2}$
0	1	2	0	3	1	14	9·892	14	4·$\dot{2}$
0	2	1	0	3	2	30	9·892	14	22·0
0	0	2	1	3	1	− 10	9·892	14	4·$\dot{2}$
2	2	1	1	6	3	− 26	35·574	50	72·6
0	0	0	2	2	2	18	4·286	6	18·0
1	2	0	0	3	2	6	9·892	14	22·0
0	2	4	1	7	3	18	47·751	66	79·$\dot{3}$
5	1	3	0	9	1	− 16	77·196	104	22·$\dot{2}$
Total 34	40	38	28	140	68	580	826·845	1,144	1,673·$\dot{7}$

For single ascertainment through affected individuals

$$k_s = (s - 1)(s + 4)$$

and is tabulated by Fisher (1935*b*).

It will be seen that $S(k_c)$ is 826·845 and $S(k_s)$ is

1,144. Therefore if we assume complete ascertainment

$$1 - 4x_c = \frac{S(u_{31})}{S(k_c)} = \frac{580}{826{\cdot}845} = 0{\cdot}7015$$

Now if segregation of achromatopsia is independent of sex, i.e. $p = 0{\cdot}5$, $1 - 4x$ *should be* 0. Hence the deviation from expectation is 0·7015. The variance of $1 - 4x$ is given by

$$V_{4x} = \frac{18}{S(k_c)} = 0{\cdot}02177$$

and $$\sigma_{4x} = \sqrt{V_{4x}} = 0{\cdot}1476$$

Hence the deviation is 4·753 times its standard error and must be considered to be highly significant (*see* Table I).

Assuming single ascertainment we find in a similar manner :

$$1 - 4x_s = \frac{S(u_{31})}{S(k_c)} = \frac{580}{1144} = 0{\cdot}5070$$

$$V_{4x} = \frac{18}{S(k_s)} = 0{\cdot}015713$$

$$\sigma_{4x} = 0{\cdot}1254$$

Again the deviation (0·5070) is 4·043 times its standard error and is highly significant.

It will be noticed that the complete ascertainment formulae give an apparently more highly significant result. It can be shown, however, that on taking an empirical test of significance, by basing the variance on the observed distribution of the families and not taking their theoretical variance, the two methods give very nearly the same significance for the deviation. This test is fully described by Fisher (1936*c*), and need not be discussed in full here.

We may next turn to a consideration of the estimate of p itself. Now

$$1-4x=1-4p(1-p)=1-4p+4p^2$$

Then $\qquad 1-2p_c=\sqrt{1-4x_c}=\sqrt{0\cdot7015}=0\cdot8376$

or $\qquad p_c=8\cdot12$ per cent

Similarly $\qquad 1-2p_s=\sqrt{0\cdot7131}$

or $\qquad p_s=14\cdot35$ per cent.

Although the two tests of significance based on $1-4x_c$ and $1-4x_s$ gave similar results when these quantities are used as the bases for estimating p they give very different answers. This is due to the different assumptions, made about the method of ascertainment, giving very different expectations for the number of recessives in the families. Hence in the absence of precise knowledge as to the actual method of ascertainment employed we may do one of two things, (*a*) take that method of ascertainment whose expectation of recessives agrees best with the observed results or (*b*) employ a method, if one can be found, independent of the method of ascertainment.

In the present case the first course is of little value as there is an excess of affected individuals even over the expectation of single ascertainment. Thus the second approach, that of finding a method independent of ascertainment, is to be preferred.

The method of doing this has been worked out by Fisher (1936). The equation of estimation is still

$1-4x_i=\dfrac{S(u_{31})}{S(k_i)}$ but now

$$k_i=\frac{1}{9}[(s_1+9s_2)^2-(s_1+81s_2)]$$

where s_1 is the number of normals in the family and s_2 the number of affected individuals in the family. The value of k_i for various s_1 and s_2 values are tabulated by Fisher (1936) and also in Table IV at the end of this book. The values of k_i for the present achromatopsia families are given in Table 39, column ten.

We then find :

$$1 - 4x_i = \frac{580}{1{,}673\cdot 7} = 0\cdot 3465$$

$$1 - 2p_i = 0\cdot 5886$$

$$p_i = 20\cdot 57 \text{ per cent.}$$

This is looser linkage than that shown by either of the other methods, as might be expected in view of the excess of affected individuals. The important point is that it is trustworthy inasmuch as it is independent of the ascertainment. In general, unless the method of ascertainment is known with exactitude, the use of k_i is preferable to the use of k_c and k_s. If the number of recessives agrees with complete or single ascertainment then k_i will give an answer closely approximating the value obtained by the use of k_c or k_s. This is well demonstrated by various examples worked by Fisher (l.c.). It must be emphasized that this k_i is applicable only to the use of u_{31}.

Returning to the general properties of u statistics it should be noted that whereas these statistics are fully efficient for the detection of linkage they are more or less inefficient for its estimation (Fisher, 1935*a*). The loss of efficiency is, however, small for recombination values above 10 per cent. and only becomes considerable for values below 5 per cent.

The formulae for the calculation of $1 - 4x$ from single backcross data are used in the above example. The corresponding formulae for the F_2 and double backcross are :

Double Backcross

$$1 - 4_x = \frac{S(u_{11})}{S(k)} \qquad V_{4x} = \frac{2}{S(k)}$$

where $$k = s(s - 1)$$

F_2

$$1 - 4x = \frac{S(u_{33})}{S(k)} \qquad V_{4x} = \frac{81}{S(k)}$$

where $$k = 2(s - 1)(s + 4)\frac{4^s - 3^{s-2}}{4^s - 3^s}$$

28. LINKAGE DETECTION WHEN THE PARENTS ARE UNKNOWN

So far we have considered the detection of linkage when both parents are known phenotypically, whatever their genotype. If but one parent is known a modified u method may be used for the analysis of linkage (Fisher, 1935*b*). But we can also detect linkage purely from a study of sibs when having no knowledge of their parents, as Penrose (1935) has shown.

Let us consider pairs of sibs from families showing segregation for two autosomal characters **A**,**a** and **B**,**b**. For character **A**,**a** the two sibs may be alike (**A** and **A** or **a** and **a**) or different (**a** and **A**). Similarly they may be alike or unlike for character **B**,**b**.

Taking the two characters together the pairs of sibs fall into four classes, as being like or unlike for **A**,**a** and like or unlike for **B**,**b**. The four classes will comprise :

TABLE 40

	A,a like	A,a unlike
B,**b** like	1. **A A** and **B B** or **a a** or **b b**	2. **A a** and **B B** or **a A** or **b b**
B,**b** unlike	3. **A A** and **B b** or **a a** or **b B**	4. **A a** and **B b** or **a A** or **b B**

The frequency of these four classes should be in simple proportion if there is no linkage, but classes 1 and 4 will be increased if linkage is in fact present. This may be tested by the calculation of χ^2 for a 2×2 contingency table as used in a number of previous examples.

If a family consists of more than two children it may be used as many times as pairs can be formed. For example, three children may be divided into and used as three pairs, four children into six pairs, &c.

Ex. 21. Penrose (l.c.) reports fifty pairs of sibs

classified for blood group B as opposed to group O and for blue eyes as opposed to not blue eyes. His results give a 2 × 2 table, of the form discussed above, thus:

TABLE 41

		B Like	B Unlike	
Blue	Like	31	2	33
	Unlike	14	3	17
		45	5	50

There is a suggestion of the classes 1 and 4 being in excess. Is this evidence for linkage significant?

It will be noticed that the expectation in two of the classes in this table is below 5 and so we cannot calculate χ^2 without correction as it would overemphasize discrepancies. This overemphasis results from the assumption of continuity in using the χ^2 distribution, whereas actually the data are discontinuous. This error can, however, be materially reduced by using Yates' (1931) Correction for Continuity which consists of reducing the two high classes each by 0·5 and similarly increasing the two low classes. On doing this the table becomes:

TABLE 42

30·5	2·5	33
14·5	2·5	17
45	5	50

χ^2 may now be calculated by the usual formula and gives

$$\frac{(30{\cdot}5 \times 2{\cdot}5 - 14{\cdot}5 \times 2{\cdot}5)2250}{45 \times 5 \times 17 \times 33} = 0{\cdot}634$$

Such a χ^2 for one degree of freedom has a probability of between 0·5 and 0·3. The suggestion of linkage is not borne out by statistical analysis.

Penrose notes that this method would, even in good circumstances, probably require nearly 100

pairs to give a significant result. Hence it should not be used unless the parents cannot be obtained or else the collection of pairs of sibs is so much easier than the collection of whole families, that vastly increased numbers of observations can be made.

The detection and estimation of linkage from human data has been extensively discussed by D. J. Finney, in a number of papers the bibliography of which can be found in the last of the series (Finney, 1942). The efficiency of the estimates and the effect of incomplete knowledge of the parents on this efficiency are considered in detail. Reference should be made to this series for a full account of the statistical methods available for dealing with human linkage.

It will be seen from this and the preceding chapter that the methods applicable to human data are related to those simpler methods in use for other genetical data. They are more complex and often less efficient than the other methods because of the shortcomings of human data themselves. These methods formulated for human data may also prove valuable in the analysis of data from other species in which the various complicating circumstances are encountered.

It will be seen from this and the preceding chapter that the methods applicable to human data are related to those simpler methods in use for other genetical data. They are more complex and often less efficient than the other methods because of the shortcomings of human data itself. These methods formulated for human data may also prove valuable in the analysis of data from other species in which the various complicating circumstances are encountered.

CHAPTER XI

THE ESTIMATION OF GENE FREQUENCIES

29. TWO ALLELOMORPHS WITH COMPLETE DOMINANCE

THE discovery of genes, such as those governing taste blindness to phenylthiocarbamide and the various blood groups for which human populations are polymorphic, has raised the problems of estimating the frequencies with which the various allelomorphs of a gene occur and of testing for differences between the gene frequencies shown by different populations, or groups within a population. The data will consist of the frequencies of the recognizable phenotypes in random samples of one or more populations; and since these samples are not restricted by family size, the methods of analysis used are the same as would be employed for other species. The special complications which arise in tests of human gene segregation and linkage are thus absent from the analysis of gene frequencies. It should be observed, however, that gene frequencies cannot be measured and analysed in populations until the mode of inheritance has been established by an adequate study of family material.

The simplest case is that of a single gene, one of whose allelomorphs is completely dominant to the other, governing the difference between two recognizable phenotypes. Ability to taste phenylthiocarbamide seems to fall into this class: individuals homozygous or heterozygous for the allelomorph **T**

taste the substance at or below the critical threshold concentration, while **tt** individuals cannot. There is no evidence of any phenotypic difference between **TT** and **Tt** individuals.

Ex. 22. Falconer (1947) records that of 629 individuals tested in the Galton Laboratory, London, 475 were tasters (i.e. **TT** or **Tt**) and 154 were non-tasters (i.e. **tt**). The relative frequencies of these two phenotypes will depend on two things, the relative frequencies of the allelomorphs **T** and **t**, and on the frequency of association of the allelomorphs in the individuals. This association, which will depend in turn on the system of mating, will determine the proportions of the phenotypically indistinguishable **TT** and **Tt** genotypes in the taster class as well as influencing the proportion of tasters and non-tasters. With, for example, close inbreeding, as occurs in wheat, all tasters would be **TT**, **Tt** individuals being absent or virtually so. There is no reason to suspect any departure from randomness of mating in respect of the gene **T**-**t** among the parents of the sample tested in London, and since some assumption must be made about the association of the allelomorphs, it will be assumed that they are randomly associated.

Let the relative frequencies of allelomorphs **T** and **t** be y and $x(= 1 - y)$ respectively. Then a sample of n individuals will be expected to contain the genotypes **TT**, **Tt** and **tt** in the numbers ny^2, $n2xy$ and nx^2. Since **TT** and **Tt** are phenotypically alike we thus expect $n(y^2 + 2xy)$ tasters and nx^2 non-tasters. If a_1 tasters and a_2 non-tasters are observed we can thus find

$$x = \sqrt{\frac{nx^2}{n}} = \sqrt{\frac{a_2}{n}}$$

as an estimate of the frequency of **t**. That of **T** is then obtained by subtraction from 1. We have already seen in Ex. 14 that such a set of data yields

$\dfrac{4n}{1-x^2}$ units of information about x, so that

$$s_x = s_y = \sqrt{V_x} = \sqrt{\frac{1-x^2}{4n}}$$

Substituting Falconer's figures in these formulae gives us

$$x = \sqrt{\frac{154}{629}} = \sqrt{0{\cdot}244833} = 0{\cdot}4948$$

and $$s_x = \sqrt{\frac{1-0{\cdot}244833}{4 \times 629}} = 0{\cdot}0173$$

Thus the frequencies of the **T** and **t** allelomorphs are 0·5052 and 0·4948 respectively, both of these having a standard error of 0·0173.

Homogeneity of the frequencies of tasters and non-tasters in different populations or groups can be tested by the χ^2 method described in Ex. 4.

30. TWO ALLELOMORPHS WITH INCOMPLETE DOMINANCE

The MN blood group system is controlled by the two allelomorphs denoted as **M** and **N**, neither of which is dominant to the other. Thus three phenotypes can be recognized, corresponding to the three genotypes, **MM**, **MN** and **NN**. The ability to recognize all three classes permits any assumption made about the association of the allelomorphs in the individuals to be tested after the gene frequencies have been calculated.

Since all three phenotypes can be recognized we can in effect detect the presence of either allelomorph irrespective of whatever its partner in the individual may be. Thus observation on n individuals may be regarded as observations on $2n$ genes. The **MM** individuals each carry two **M** genes and the **MN** individuals one **M** gene. Thus where a_1, a_2, and a_3 are the numbers observed to fall into the **MM**, **MN**

and NN categories respectively, we can estimate the relative frequency of allelomorph M as $x = \frac{2a_1 + a_2}{2n}$ and that of allelomorph N as $y = 1 - x = \frac{a_2 + 2a_3}{2n}$. It is easy to show that these are maximum likelihood estimates, and that they are independent of the association of the allelomorphs in the genotypes.

Since we have partitioned $2n$ genes into the two classes with proportions x and y, the standard error of x and also of y will be given by

$$s_x = s_y = \sqrt{\frac{xy}{2n}}$$

This again can be verified by the methods of Chapters V and VI.

Where data are available from a number of populations, their homogeneity in respect of x and y can be tested by making a table listing the values of $2a_1 + a_2$ and $a_2 + 2a_3$ for each population and calculating a homogeneity χ^2 by Brandt and Snedecor's method as described in Ex. 4. It is, however, informative to look at the analysis of χ^2 from such data in another way.

The frequencies of three classes are available from any sample of individuals that has been tested. Assuming random mating and random association their expected proportions are as set out in Table 43.

TABLE 43

Class	MM	MN	NN	Total
Number observed .	a_1	a_2	a_3	n
Proportion expected .	x^2	$2xy$	y^2	1

A χ^2 for 2 degrees of freedom can be calculated to test agreement with the expectations derived by using

any given values of x and y. This χ^2 can be partitioned into two parts, each corresponding to a single degree of freedom, by the use of orthogonal functions. One of these parts can usefully be made to test the agreement of the data with the particular values of x and y used. From Section 8, the appropriate function will clearly be $y(2a_1 + a_2) - x(a_2 + 2a_3)$ whose variance can be shown (Section 12) to be $2xyn$. The test of agreement with x and y will thus be given by

$$\chi^2_F = \frac{[y(2a_1 + a_2) - x(a_2 + 2a_3)]^2}{2xyn}$$

The second simple χ^2, orthogonal to this first part, can be found by subtraction from the compound χ^2 as

$$\chi^2_M = \frac{[y^2a_1 - xya_2 + x^2a_3]^2}{x^2y^2n}$$

The function $(y^2a_1 - xya_2 + x^2a_3)$, upon which this depends, compares the frequency of heterozygotes, MN, with those of the two homozygous classes. In other words it is a test of the randomness of association of the allelomorphs in the genotypes, i.e. of the randomness of mating, other things being equal. Strictly, it is a valid test only when the data do not disagree with the values of x and y used. When the best-fitting values of $x = \frac{2a_1 + a_2}{2n}$ and $y = \frac{a_2 + 2a_3}{2n}$ are used, this formula for χ^2_M is identical with $\chi^2 = \frac{n(a_2{}^2 - 4a_1a_3)^2}{(2a_1 + a_2)^2(a_2 + 2a_3)^2}$ testing the agreement with the expectation $a_2{}^2 = 4a_1a_3$ which is the expectation of random association, for then it is equivalent to $(2nxy)^2 = 4nx^2ny^2$.

Ex. 23. Both the use and the limitations of this type of analysis can be illustrated from data collected by Landsteiner and Levine from Whites, Negroes and Amerinds of the U.S.A. (quoted by Taylor and Prior, 1939). They found the following frequencies :—.

TABLE 44

Class	MM	MN	NN	Total	χ^2_F	χ^2_M
Whites .	139	285	108	532	11·8285	2·2986
Negroes .	50	86	45	181	6·7444	0·8596
Amerinds .	123	72	10	205	63·6906	4·5445
Total . .	312	443	163	918	82·2635	7·7027

Pooling the figures from the three groups we find

$$x = \frac{(2 \times 312) + 443}{2 \times 918} = 0{\cdot}581155$$

and $\quad y = 0{\cdot}418845$

$$s_x = s_y = \sqrt{\frac{0{\cdot}581155 \times 0{\cdot}418845}{2 \times 918}} = 0{\cdot}011514$$

These estimates of x and y may then be used in the tests of homogeneity and randomness of mating by the χ^2 developed above.

The values of χ^2_F testing agreement with the pooled values of x and y, and χ^2_M testing randomness of association, are given in the sixth and seventh columns of Table 44. The two χ^2 for the whites are calculated as

$$\chi^2_F = \frac{[0{\cdot}418845(2 \times 139 + 285) - 0{\cdot}581155(285 + 2 \times 108)]^2}{2 \times 532 \times 0{\cdot}581155 \times 0{\cdot}418845} = 11{\cdot}8285$$

and

$$\chi^2_M = \frac{[(0{\cdot}418845^2 \times 139) - (0{\cdot}418845 \times 0{\cdot}581155 \times 285) + (0{\cdot}581155^2 \times 108)]^2}{0{\cdot}581155^2 \times 0{\cdot}418845^2 \times 532}$$

$$= 2{\cdot}2986$$

and the rest similarly.

Summing χ^2_F over the three groups, we obtain $S(\chi^2_F) = 82{\cdot}2635$ for two degrees of freedom since of the original three, one from each group, one has been

used up in estimating x and y from the pooled data. This χ^2 is clearly very significant, and the groups cannot be regarded as homogeneous for the gene frequencies. Inspection shows that this is due to the frequency of allelomorph M being much higher in the Amerinds than in Whites and Negroes who are much alike.

Summing, χ^2_M gives $S(\chi^2_M) = 7{\cdot}7027$ for, of course, three degrees of freedom, since none has been used up in making any estimates. This has a probability of just over 0·05 so that, while there is no clear evidence of non-random mating it might be supposed that there was at least some suspicion of it.

We have just seen, however, that the groups are inhomogeneous in x and so the values obtained for χ^2_M using a common estimate of x must themselves be under suspicion. If, instead of the common estimate of x the estimates obtained from the individual groups are used, we find

Whites . .	$x = 0{\cdot}529135$,	$\chi^2_M = 2{\cdot}9988$
Negroes . .	$x = 0{\cdot}513812$,	$\chi^2_M = 0{\cdot}4345$
Amerinds . .	$x = 0{\cdot}775610$,	$\chi^2_M = 0{\cdot}0167$

and there is no real ground for suspecting departure from random mating. The Amerinds who gave the highest χ^2_M when the common value of x was used now give the lowest. This group is one whose individual value of x departs most from the common or pooled value, and so is the one whose χ^2_M calculated using the common value would be expected to be the least trustworthy.

31. MORE THAN TWO ALLELOMORPHS

Estimation of the gene frequencies is simple where only two allelomorphs are recognized. With three or more allelomorphs, estimation may be more troublesome. The use of the method of maximum likelihood in such a complex case may be illustrated by reference to the A—B—O system of blood groups in man.

Three allelomorphs of this gene are normally recognized, viz. **O**, **A** and **B**. It is known that there are in fact more allelomorphs, in particular that **A** is a compound group including $\mathbf{A}_1$ and $\mathbf{A}_2$, but this further subdivision will be neglected for our purpose. The method used in the analysis of the more complex case is a simple and obvious extension of that which will be illustrated.

Now **O** can be inferred only by the absence of both **A** and **B**, but **A** and **B** can be detected even in one another's presence. Thus four phenotypes are recognizable as in Table 45.

Let the gene frequencies be x, y and $z(= 1 - x - y)$ for **A**, **B** and **O** respectively. Then the four phenotypes are expected with the frequencies shown in the table.

TABLE 45

Phenotype	Genotypes	Proportion Expected	Number Observed
O	**OO**	z^2	55
A	**AA** and **AO**	$x^2 + 2xz$	69
B	**BB** and **BO**	$y^2 + 2yz$	33
AB	**AB**	$2xy$	15

The last column of the table gives the numbers falling into the four phenotypic classes out of 172 Latvian's recorded by Race *et al.* (1948).

The log likelihood expression for the estimation of x, y and z is thus

$$L = 55 \log z^2 + 69 \log (x^2 + 2xz) + 33 \log (y^2 + 2yz) + 15 \log 2xy$$

$$= 110 \log z + 84 \log x + 69 \log (x + 2z) + 48 \log y + 33 \log (y + 2z) + 15 \log 2$$

This expression appears to involve three unknowns, x, y and z, but actually there are only two, for we have the additional condition that $x + y + z = 1$.

Thus we can substitute $1 - x - y$ for z, this being more convenient than the possible alternative substitutions for x and y.

In order to estimate x and y simultaneously the log likelihood must be partially differentiated with respect to x and y in turn to yield two equations of estimation. In so doing we need not first substitute for z since $\frac{\partial}{\partial x}$ with substitution $= \frac{\partial}{\partial x} - \frac{\partial}{\partial z}$ without substitution. Then partial differentiation gives as the equation of estimation

$$\frac{\partial L}{\partial x} = -\frac{110}{z} + \frac{84}{x} - \frac{69}{x + 2z} - \frac{66}{y + 2z} = 0$$

$$\frac{\partial L}{\partial y} = -\frac{110}{z} - \frac{138}{x + 2z} + \frac{48}{y} - \frac{33}{y + 2z} = 0$$

These equations cannot easily be solved algebraically. There is, however, a method, which has been used for the estimation of linkage values by Mather (1935) and later for the estimation of gene frequencies by Stevens (1938), of substituting approximate trial values for x and y and calculating efficient correction for them.

Let us take as trial values

$$z = \sqrt{\frac{55}{172}} = 0{\cdot}56547983$$

$$x = \sqrt{\frac{55 + 69}{172}} - z = 0{\cdot}28359628$$

$$y = 1 - x - z = 0{\cdot}15092389$$

Then substituting these values in the left sides of the two equations of estimation gives $\frac{\partial L}{\partial x} = 1{\cdot}405346$ and $\frac{\partial L}{\partial y} = 0{\cdot}215551$. The fact that these expressions depart numerically from 0 shows, of course, that the trial values of x, y and z are not maximum likelihood estimates.

The next step is to calculate three quantities defined as

$$I_{xx} = nS\left[\frac{1}{m}\left(\frac{\partial m}{\partial x}\right)^2\right]$$

$$I_{yy} = nS\left[\frac{1}{m}\left(\frac{\partial m}{\partial y}\right)^2\right]$$

$$I_{xy} = nS\left[\frac{1}{m}\left(\frac{\partial m}{\partial x}\right)\left(\frac{\partial m}{\partial y}\right)\right]$$

The relation borne by these quantities to the amounts of information discussed in Chapter VI requires no comment. Now the amount of information about x is, as we have seen, the rate of change of $\frac{dL}{dx}$ on x when x is near its trial value. Similarly, I_{xx}, &c., give the rates of change of $\frac{\partial L}{\partial x}$ and $\frac{\partial L}{\partial y}$ on x and y so that we can estimate δx and δy, the corrections needed to convert the trial values of x and y into maximum likelihood estimates, from the equations

$$I_{xx}\delta x + I_{xy}\delta y = 1{\cdot}405346$$

$$I_{xy}\delta x + I_{yy}\delta y = 0{\cdot}215551$$

The calculation of I_{xx}, &c., is set out in Table 46.

TABLE 46

Class	m	$\frac{\partial m}{\partial x}$	$\frac{\partial m}{\partial y}$	I_{xx}	I_{xy}	I_{yy}
O . .	z^2	$-2z$	$-2z$	688·0000	688·0000	688·0000
A . .	$x^2 + 2xz$	$2z$	$-2x$	548·4058	$-$ 275·0334	137·9332
B . .	$y^2 + 2yz$	$-2y$	$2z$	81·0024	$-$ 303·4988	1137·1457
AB .	$2xy$	$2y$	$2x$	183·0695	344·0000	646·3995
Total .	1	0	0	1500·4777	453·4678	2609·4784

Thus, for example, the contribution of Class O to

I_{xx} is $\frac{4z^2}{z^2}n = 4 \times 172 = 688$ and that of Class A to

$I_{xy} = -\frac{4xz}{x^2 + 2xz}n = -275{\cdot}0334$, the trial values of x, y and z being substituted in the calculations.

Then the equations become

$$1500{\cdot}4777\delta x + 453{\cdot}4678\delta y = 1{\cdot}405346$$
$$453{\cdot}4678\delta x + 2609{\cdot}4784\delta y = 0{\cdot}215551$$

and $\delta x = 0{\cdot}000962, \quad \delta y = -0{\cdot}000084$

so that the efficient estimate of

$$x = x + \delta x = 0{\cdot}283596 + 0{\cdot}000962 = 0{\cdot}28456$$

and of

$$y = y + \delta y = 0{\cdot}150923 - 0{\cdot}000084 = 0{\cdot}15084$$

leaving

$$z = 1 - x - y = 0{\cdot}56460$$

Since I_{xx}, &c. were estimated using the trial values of x, y and z, they did not have their maximum likelihood values in this calculation. The whole process may therefore be repeated using the new estimates as trial values and finding new corrections. Provided, however, trial values were chosen reasonably near the true values, as in the present case, a second round of calculations will not be necessary.

One thing remains, viz. to find the standard errors of x, y and z. Now the variance of x can be found as

$$V_x = 1 \bigg/ \left(I_{xx} - \frac{I^2_{xy}}{I_{yy}}\right)$$

and similarly $V_y = 1 \bigg/ \left(I_{yy} - \frac{I^2_{xy}}{I_{xx}}\right)$

Now the estimates of x and y must obviously be correlated and their covariance

$$W_{xy} = 1 \bigg/ \left(I_{xy} - \frac{I_{xx}I_{yy}}{I_{xy}}\right)$$

Since $z = 1 - x - y$

$$V_z = V_x + V_y + 2W_{xy}$$

$$= \frac{I_{xx} + I_{yy} - 2I_{xy}}{I_{xx}I_{yy} - I^2_{xy}}$$

Then substituting for I_{xx}, &c. the values found in Table 46 give

$$V_x = 1 \Big/ \left(1500{\cdot}4777 - \frac{453{\cdot}4678^2}{2609{\cdot}4784}\right) = 0{\cdot}000703395$$

$$V_y = \qquad 0{\cdot}000404460$$

$$V_z = \qquad 0{\cdot}000863387$$

giving

$$s_x = \sqrt{V_x} = 0{\cdot}026522$$

$$s_y = \sqrt{V_y} = 0{\cdot}020111$$

$$s_z = \sqrt{V_z} = 0{\cdot}029383$$

It will be observed that δx was only about $\frac{1}{25}s_x$ and δy was even smaller in relation to s_y. The trial estimates were indeed so well chosen that the necessary corrections were quite trivial.

CHAPTER XII

SYMBOLS AND FORMULAE

Symbols

a number observed in a class.

A_1 sum of a_1 and a_4 in a four-class segregation.

A_2 sum of a_2 and a_3 in a four-class segregation.

D_x deviation from zero of maximum likelihood expression of x.

f misclassification due to incomplete manifestation of a character.

i_x $= \frac{I_x}{n} = \frac{n}{V_x}$ *amount of information concerning* x per individual in a family.

k coefficient in orthogonal functions.

l the characteristic proportion in a two-class segregation which may be represented as $l : 1$.

$l = \frac{x}{1 - x}$.

m proportion expected in any class.

n number of individuals in a family.

p recombination fraction.

P $= p^2$ or $(1 - p)^2$ in F_2 data.

s_x standard error of x ($\sigma_x = s_x$ when x is **fixed by** hypothesis).

S summation over all classes.

V_x variance of $x = (s_x)^2 = \frac{1}{ni_x}$.

x (i) $1 - y =$ chance of any individual being of a chosen type in a two-class segregation.

(ii) Also used as $p(1 - p)$ in human data.

(iii) Also used in the analysis of χ^2 by orthogonal functions.

y proportion of recessives as found by the sib method.

Binomial Expansion

$$(x + y)^n = x^n + nx^{n-1}y + {}^nC_2x^{n-2}y^2 \ldots nxy^{n-1} + y^n$$

$$\sigma_x = \sqrt{\frac{xy}{n}}$$

$$\sigma_{nx} = \sqrt{xyn}$$

χ^2

$$\chi^2 = S\left[\frac{(a - nm)^2}{nm}\right] = S\left(\frac{a^2}{nm}\right) - n$$

for a two-class segregation expected to be $l : 1$

$$\chi^2 = \frac{(a_1 - la_2)^2}{ln}$$

Brandt and Snedecor formula for testing heterogeneity

$$\chi^2 = \frac{n_t^2}{a_{1t}a_{2t}}\left[S\left(\frac{a_1^2}{n}\right) - \frac{a_{1t}^2}{n_t}\right]$$

From 2×2 contingency table

$$\chi^2 = \frac{(a_1a_4 - a_2a_3)^2n}{(a_1 + a_2)(a_3 + a_4)(a_2 + a_4)(a_1 + a_3)}$$

For the detection of linkage between two factors segregating into $l_1 : 1$ and $l_2 : 1$ respectively

$$\chi^2 = \frac{(a_1 - l_2a_2 - l_1a_3 + l_1l_2a_4)}{l_1l_2n}$$

Estimation

Likelihood expression is

$$\frac{n!}{a_1!a_2! \ldots} (m_1)^{a_1}(m_2)^{a_2} \ldots$$

Logarithm likelihood

$$L = C + a_1 \log m_1 + a_2 \log m_2 \ldots$$

Equation of estimation by maximum likelihood

$$\frac{dL}{dp} = a_1\frac{d \log m_1}{dp} + a_2\frac{d \log m_2}{dp} \ . \ . \ . = 0$$

$$i_p = \frac{I_p}{n} = \frac{n}{V_p} = -S\left(m\frac{d^2 \log m}{dp^2}\right) = S\left[\frac{1}{m}\left(\frac{dm}{dp}\right)^2\right]$$

$$i_p = i_x\left(\frac{dx}{dp}\right)^2$$

$$\chi^2 = S\left(\frac{{D_x}^2}{I_x}\right)$$

to test heterogeneity between bodies of data.

Product equation of estimation

$$\frac{a_1 a_4}{a_2 a_3} = \frac{m_1 m_4}{m_2 m_3}$$ [fully efficient for linkage estimation]

Human Data

$$V_y = \frac{1}{n'}\frac{y(1-y)}{s-1}(1 + y' + 2yy') \text{ where}$$

$$y' = \frac{S\{t(t-1)n_{at}\}}{S\{t(9-1)n_{at}\}}$$

$$u_{11} = (a_1 - a_2 - a_3 + a_4)^2 - (a_1 + a_2 + a_3 + a_4)$$
$$u_{31} = (a_1 - a_2 - 3a_3 + 3a_4)^2 - (a_1 + a_2 + 9a_3 + 9a_4)$$
$$u_{33} = (a_1 - 3a_2 - 3a_3 + 9a_4)^2 - (a_1 + 9a_2 + 9a_3 + 81a_4)$$

$$1 - 4x = \frac{S(u)}{S(k)}$$ (k depends on ascertainment, page 113 *et seq.*)

$$V_{(1-4x)} = \frac{2}{S(k)} \text{ for } u_{11}$$

$$= \frac{18}{S(k)} \text{ for } u_{31}$$

$$= \frac{81}{S(k)} \text{ for } u_{33}$$

Special Formulae

1 : 1 ratio $\sigma_{nx} = \frac{1}{2}\sqrt{n}$ $\quad \chi^2 = \dfrac{(a_1 - a_2)^2}{n}$

3 : 1 ratio $\sigma_{nx} = \frac{1}{4}\sqrt{3n}$ $\quad \chi^2 = \dfrac{(a_1 - 3a_2)^2}{3n}$

15 : 1 ratio $\sigma_{nx} = \frac{1}{16}\sqrt{15n}$ $\quad \chi^2 = \dfrac{(a_1 - 15a_2)^2}{15n}$

9 : 7 ratio $\sigma_{nx} = \frac{1}{16}\sqrt{63n}$ $\quad \chi^2 = \dfrac{(a_1 - \frac{9}{7}a_2)^2}{\frac{9}{7}n}$

Linkage between factors segregating in :

1 : 1 and 1 : 1 (Backcross)—

$$\chi^2 = \frac{(a_1 - a_2 - a_3 + a_4)^2}{n}$$

$$p = \frac{a_2 + a_3}{n} \quad V_p = \frac{p(1 - p)}{n}$$

1 : 1 and 3 : 1 (Single backcross)—

$$\chi^2 = \frac{(a_1 - a_2 - 3a_3 + 3a_4)^2}{3n}$$

p given by solution of

$$-\frac{a_1}{2 - p} + \frac{a_2}{1 + p} + \frac{a_3}{p} - \frac{a_4}{1 - p} = 0$$

$$V_p = \frac{2p(1 - p)(1 + p)(2 - p)}{n(1 + 2p - 2p^2)}$$

3 : 1 and 3 : 1 (F_2)

$$\chi^2 = \frac{(a_1 - 3a_2 - 3a_3 + 9a_4)^2}{9n}$$

$P = p^2$ or $(1 - p)^2$ is given by the solution of

$$nP^2 - (a_1 - 2a_2 - 2a_3 - a_4)P - 2a_4 = 0$$

$$V_P = \frac{2P(1 - P)(2 + P)}{n(1 + 2P)}$$

and $V_p = \dfrac{(1 - P)(2 + P)}{2n(1 + 2P)}$ where $P = p^2$ or $(1 - p)^2$

Product Formulae in F_2

P is the solution of

$$(a_1a_4 - a_2a_3)P^2 - 2(a_1a_4 + a_2a_3)P + a_1a_4 = 0$$

$$V_P = \frac{2P(1 - P)(2 + P)}{n(1 + 2P)}$$

TABLE Ia (Fisher 1936a)

Table of Normal Deviates

The deviation in the normal distribution in terms of the standard deviation

	0·01	0·02	0·03	0·04	0·05	0·06	0·07	0·08	0·09	0·10
0·00	2·575829	2·326348	2·170090	2·053749	1·959964	1·880794	1·811911	1·750686	1·695398	1·644854
0·10	1·598193	1·554774	1·514102	1·475791	1·439521	1·405072	1·372204	1·340755	1·310579	1·281552
0·20	1·253565	1·226528	1·200359	1·174987	1·150349	1·126391	1·103063	1·080319	1·058122	1·036433
0·30	1·015222	0·994458	0·974114	0·954165	0·934589	0·915365	0·896473	0·877896	0·859617	0·841621
0·40	0·823894	0·806421	0·789192	0·772193	0·755415	0·738847	0·722479	0·706303	0·690309	0·674490
0·50	0·658838	0·643345	0·628006	0·612813	0·597760	0·582841	0·568051	0·553385	0·538836	0·524401
0·60	0·510073	0·495850	0·481727	0·467699	0·453762	0·439913	0·426148	0·412463	0·398855	0·385320
0·70	0·371856	0·358459	0·345125	0·331853	0·318639	0·305481	0·292375	0·279319	0·266311	0·253347
0·80	0·240426	0·227545	0·214702	0·201893	0·189118	0·176374	0·163658	0·150969	0·138304	0·125661
0·90	0·113039	0·100434	0·087845	0·075270	0·062707	0·050154	0·037608	0·025069	0·012533	0

The value of P for each entry is found by adding the column heading to the value in the left-hand margin. The corresponding value of $\frac{d}{\sigma}$ is the deviation such that the probability of an observation falling outside the range from $-\frac{d}{\sigma}$ to $+\frac{d}{\sigma}$ is P. For example, $P = 0{\cdot}03$ for $\frac{d}{\sigma} = 2{\cdot}170090$; so that 3 per cent of normally distributed values will have positive or negative deviations exceeding the standard deviation in the ratio 2·170090 at least.

TABLE Ib

Values of $\frac{d}{\sigma}$ for Small Values of P

P . . .	0·001	0·000,1	0·000,01	0·000,001	0·000,000,1	0·000,000,01	0·000,000,001
$\frac{d}{\sigma}$ . . .	3·29053	3·89059	4·41717	4·89164	5·32672	5·73073	6·10941

TABLE II—Table of χ^2 (Fisher 1936*a*)

n	P = 0·99	0·98	0·95	0·90	0·80	0·70	0·50	0·30	0·20	0·10	0·05	0·02	0·01
1	0·000157	0·000628	0·00393	0·0158	0·0642	0·148	0·455	1·074	1·642	2·706	3·841	5·412	6·635
2	0·0201	0·0404	0·103	0·211	0·446	0·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210
3	0·115	0·185	0·352	0·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·341
4	0·297	0·429	0·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277
5	0·554	0·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086
6	0·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892

TABLE III

Fraction expected	Level of Probability 0·900	0·950	0·980	0·990	0·995	0·998	0·999
$\frac{1}{2}$	3·3	4·3	5·6	6·6	7·6	9·0	10·0
$\frac{1}{4}$	8·0	10·4	13·6	16·0	18·4	21·6	24·0
$\frac{1}{8}$	17·2	22·4	29·3	34·5	39·7	46·5	51·7
$\frac{1}{16}$	35·7	46·4	60·6	71·4	82·1	96·3	107·0
$\frac{1}{32}$	72·5	94·4	123·2	145·1	166·9	195·7	217·6
$\frac{1}{64}$	146·2	190·2	248·4	292·4	336·4	394·6	438·6
$\frac{1}{3}$	5·7	7·4	9·6	11·4	13·1	15·3	17·0
$\frac{1}{9}$	19·5	25·4	33·2	39·1	45·0	52·8	58·6
$\frac{1}{27}$	61·0	79·4	103·6	122·0	140·4	164·7	183·0

The numbers in the body of the table are the numbers of individuals which should be raised, in a progeny, in order that a certain type, expected to form a known fraction of the progeny, may be expected to occur, with a chosen level of probability, at least once. For example, suppose on selfing a plant heterozygous for one gene (i.e. **Aa**) we want to raise a family sufficiently large to contain at least one recessive (**aa**) in 99 cases out of 100. Recessives types are expected in $\frac{1}{4}$ of the cases. Then taking the second row of the table (the $\frac{1}{4}$ row) and the fourth column (probability 0·99) we find that sixteen plants are needed.

TABLE IV (Fisher, 1936)

TABLE OF ki $[= \frac{1}{9}(s_1 + 9s_2)^2 - (s_1 + 81s_2)]$ FOR USE WITH u_{31}

s_1 = number of normal children
s_2 = number of affected children

	s_2 1	2	3	4	5	6	7	8	9
s_1 0	—	18·0	54·0	108·0	180·0	270·0	378·0	504·0	648·0
1	2·0	22·0	60·0	116·0	190·0	282·0	392·0	520·0	
2	4·2̇	26·2̇	66·2̇	124·2̇	200·2̇	294·2̇	406·2̇	536·2̇	
3	6·6̇	30·6̇	72·6̇	132·6̇	210·6̇	306·6̇	420·6̇	552·6̇	
4	9·3̇	35·3̇	79·3̇	141·3̇	221·3̇	319·3̇	435·3̇		
5	12·2̇	40·2̇	86·2̇	150·2̇	232·2̇	332·2̇	450·2̇		
6	15·3̇	45·3̇	93·3̇	159·3̇	243·3̇	345·3̇	465·3̇		
7	18·6̇	50·6̇	100·6̇	168·6̇	254·6̇	358·6̇			
8	22·2̇	56·2̇	108·2̇	178·2̇	266·2̇	372·2̇			
9	26·0	62·0	116·0	188·0	278·0	386·0			
10	30·0	68·0	124·0	198·0	290·0				
11	34·2̇	74·2̇	132·2̇	208·2̇	302·2̇				
12	38·6̇	80·6̇	140·6̇	218·6̇	314·6̇				
13	43·3̇	87·3̇	149·3̇	229·3̇					
14	48·2̇	94·2̇	158·2̇	240·2̇					
15	53·3̇	101·3̇	167·3̇	251·3̇					
16	58·6̇	108·6̇	176·6̇						

REFERENCES

BAILEY, N. T. J. (1949.) The estimation of linkage with differential viability, II and III. *Heredity*, **3**, 220–8.

BATESON, W. (1909.) *Mendel's Principles of Heredity.* University Press, Cambridge.

FALCONER, D. S. (1947.) Sensory thresholds for solutions of phenyl-thio-carbamide. *Ann. Eugenics*, **13**, 211–22.

FINNEY, D. J. (1942.) The detection of linkage. VI. The loss of information from incompleteness of parental records. *Ann. Eugenics*, **11**, 233–44.

FISHER, R. A. (1921.) On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.*, A, **122**, 309–68.

—— (1928.) On a property connecting the χ^2 measure of discrepancy with the method of maximum likelihood. *Atti del Congresso Internationale dei Matimatici (Bologna)*, **6**, 95–100.

—— (1934.) The effect of methods of ascertainment upon the estimation of frequencies. *Ann. Eugenics*, **6**, 13–25.

—— (1935*a*.) The detection of linkage with 'Dominant' abnormalities. *Ann. Eugenics*, **6**, 187–201.

—— (1935*b*.) The detection of linkage with 'Recessive' abnormalities. *Ann. Eugenics*, **6**, 339–51.

—— (1936.) Tests of significance applied to Haldane's data on partial sex-linkage. *Ann. Eugenics*, **7**, 87–104.

—— (1946.) A system of scoring linkage data, with special reference to the pied factors in mice. *Amer. Nat.*, **80**, 568–78.

—— (1949*a*.) *Statistical Methods for Research Workers.* (11th edn.) Oliver and Boyd, Edinburgh.

—— (1949*b*.) *The Design of Experiments.* (5th edn.) Oliver and Boyd, Edinburgh.

—— and BALMAKUND, B. (1928.) The estimation of linkage from the offspring of selfed heterozygotes. *Jour. Genet.*, **20**, 79–92.

—— and MATHER, K. (1936.) A linkage test with mice. *Ann. Eugenics*, **7**, 265–80.

HALDANE, J. B. S. (1936.) A search for incomplete sex-linkage in man. *Ann. Eugenics*, **6**, 339–51.

HUTCHINSON, J. B. (1929.) The application of the 'Method of Maximum Likelihood' to the estimation of linkage. *Genetics*, **14**, 519–37.

IMAI, Y. (1931.) Linkage studies in *Pharbitis Nil* I. *Genetics*, **16**, 26–41.

IMMER, F. R. (1930.) Formulae and tables for calculating linkage intensities. *Genetics*, **15**, 81–98.

—— (1934.) Calculating linkage intensities from F_3 data. *Genetics*, **19**, 119–36.

JENKINS, M. T. (1927.) A factor for yellow-green chlorophyll colour in maize and its linkage relations. *Genetics*, **12**, 498–518.

MATHER, K. (1935.) The combination of data. *Ann. Eugenics*, **6**, 399–410.

—— (1936*a*.) Types of linkage data and their value. *Ann. Eugenics*, **7**, 251–64.

—— (1936*b*.) Segregation and linkage in autotetraploids. *Jour. Genet.*, **32**, 287–314.

—— (1937.) The analysis of single factor segregations. *Ann. Eugenics*, **8**, 96–105.

NABOURS, R. K., LARSON, I., and HARTWIG, N. (1933.) Inheritance of colour patterns in the grouse locust, *Acrydium arenosum*, Burmeister (Tettigidae). *Genetics*, **18**, 159–72.

PEARSON, K., NETTLESHIP, E., and USHER, C. H. (1913.) A monograph on albinism in man IV. *Drapers Company Research Memoirs, Biometrica*, IX.

PENROSE, L. S. (1935.) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann. Eugenics*, **6**, 133–8.

PHILP, J. (1934.) The genetics of *Papaver Rhoeas* and related forms. *Jour. Genet.*, **28**, 175–204.

RACE, R. R., SANGER, R., LAWLER, S., and KEETCH, D. V. (1948.) Blood groups of Latvians. *Ann. Eugenics*, **14**, 134–8.

SMITH, H. F. (1937.) A test of significance of Mendelian ratios when classification is uncertain. *Ann. Eugenics*, **8**, 94–5.

STEVENS, W. L. (1938.) Estimation of blood group gene frequencies. *Ann. Eugenics*, **8**, 362–75.

TAYLOR, G. L., and PRIOR, A. M. (1939.) The distribution of the M and N factors in random samples of different races. *Ann. Eugenics*, **9**, 97–108.

WINTON, D. DE, and HALDANE, J. B. S. (1935.) The genetics of *Primula sinensis*. III. Linkage in the diploid. *Jour. Genet.*, **31**, 67–100.

YATES, F. (1931.) Contingency tables involving small numbers and the χ^2 test. *Supp. Jour. Roy. Stat. Soc.*, **1**, 217–35.

INDEX

Achromatopsia, 117
Acridium arenosum, 97
Albinism, 109
Antirrhinum majus, 14
Ascertainment, 106, 117
Autotetraploids, *see Simplex*

Backcross, 3, 8, 43, 45, 79, 116, 121
— exact treatment, 96 *et seq.*
— singles, 78, 116 *et seq.*
Bailey, 96
Balmakund, 54
Bateson, 73, 87
Binomial distribution, 8
Blood groups, 125
MN—, 127 *et seq.*
ABO—, 131 *et seq.*
Brandt, 22, 38, 128

χ^2, 10, 15 *et seq.*, 51, 53, 77, 90, 93, 105, 111, 127, 129 *et seq.*
Classification, 63
Comparisons, independent, 34, 38
Complementary factors, 3, 64, 92
Consistency, criterion of, 44
Contingency table, 92, 122
Continuity, correction for, 123
Continuous distribution, 10
Coupling, 4, 61, 96, 117
Covariance of estimates, 135

Degrees of freedom, 16, 33
— loss of, 22, 54
Deviation, 69, 77
Diploid, 2
Dominance, 60, 114
Drosophila melanogaster, 2
Duplicate factors, 41

Efficiency, criterion of, 45
— of statistics, 51 *et seq.*, 96, 121
Estimation, combined, 69 *et seq.*
— criteria of, 44 *et seq.*
— equation of, 46
— inefficient, 54
— method of, 7, 11, 44
— scoring method of, 86 *et seq.*
— simultaneous, 133 *et seq.*
Eye colour, 123

F_2, 3, 36, 45, 57, 79, 88, 116, 121
— disturbed, 91, 96
Falconer, 126
Family size, 26 *et seq.*, 105 *et seq.*
Finney, 124
Fisher, 17, 24, 45, 51, 54, 86, 96, 99 *et seq.*, 116 *et seq.*
Frequency distribution, 8

Gene frequencies, 125 *et seq.*
Grouse locusts, *see Acridium*

Haldane, 52, 114
Heterogeneity, of data, 17 *et seq.*, 37, 73 *et seq.*

Heterozygote, 3, 26
Hierarchical tables, 20
Homozygote, 26
Human data, 104
Hutchinson, 64

Imai, 36, 48
Immer, 54, 55
Information, 53, 72, 77, 87, 111, 134
— calculation of, 57 *et seq.*
Interpolation, solution by, 71

Jenkins, 78, 92

Landsteiner and Levine, 129
Latvians, 132
Likelihood, 46, 70
— logarithm of, 46, 70
Linear functions, 39
Linkage, 4, 7, 44
— with sex, 114

Maize, 78, 92
Manifestation, of characters, 84, 95
Mather, 17, 57 *et seq.*, 69 *et seq.*
Matings, human, 105
Maximization, 46, 70
Maximum Likelihood, method of, 11, 45 *et seq.*, 77, 86, 91, 95, 131
Mean information, 73
— weighted, 111
Mice, 17, 22
Misclassification, 27, 29
Multinomial distribution, 11

Nabours, 97
Nettleship, 109
Normal deviates, 14, 74
— distribution, 10

Orthogonal functions, 32 *et seq.*, 38 *et seq.*, 129
Orthogonal functions, derivation of, 40
Ovule sterility, 113

Parameter, 7, 10, 44
Pearson, 15, 109
Penrose, 121
Pharbitis, 36, 48
Phenylthiocarbamide, **125** *et seq.*
Philp, 34, 47
Pisum, 113
Precision, 12, 28
Primula sinensis, 2, 52
Proband method, 106
Product formula, 54, **94**,
— tables of, 55
Poppy, 34, 47

Race *et al.*, 132
Random mating, 126
— test of, 129
Recessives, human, 116
Recombination, 4, 5, 44
Repulsion, 4, 61, 96, 117
Retinitis pigmentosum, 114

Sansome, 69
Score, 87
Segregation, 2, 29, 69
Separation, index of, 69
Sib method, 106 *et seq.*
Significance, test of, 5, 9
— meaning of, 6
Simplex autotetraploid, 69
Single factors, 2, 13
— in humans, 104
Smith, 84
Snedecor, 22, 38, 128
Standard error, 10, 13, 48, 50, 73, 109, 119
Statistic, 10, 11
Sufficiency, criterion of, 45
Sweet Pea, 73, 87

Taste blindness, 125 *et seq.*
Tetrad analysis, 3, 4
Tomato, 70